

**QUEUE DISTRIBUTION OF REAL TIME
TRANSPORTATION OF
VOICE OVER IP**

By

RITU SINGH

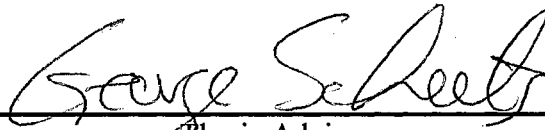
Bachelor of Science
The University of Tulsa
Tulsa, Oklahoma
1990

Master of Science
The University of Tulsa
Tulsa, Oklahoma
1996

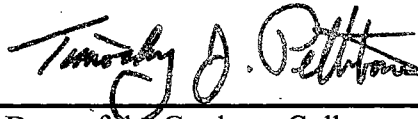
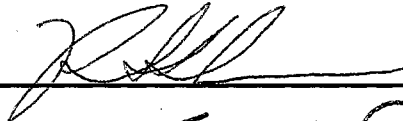
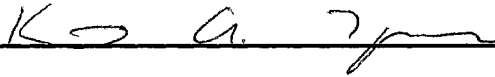
Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the degree of
DOCTOR OF PHILOSOPHY
August, 2002

**QUEUE DISTRIBUTION OF REAL TIME
TRANSPORTATION OF
VOICE OVER IP**

Thesis Approved:



Thesis Advisor



Dean of the Graduate College

ACKNOWLEDGMENTS

I wish to express my sincere appreciation to Dr. George Scheets for his supervision and guidance. I would like to thank the committee members, Dr. Keith Teague, Dr. Jong-Moon Chung and Dr. Rick Wilson for serving on my committee and providing suggestions for improvement in this work. I also wish to express my gratitude to Williams Communications Group for providing financial assistance for this study.

I would like to extend a special thanks to Navin, Davin, Surendra, and my parents, Rattan Singh and Shakuntla Sandhu, for their encouragement and support.

TABLE OF CONTENTS

Chapter	Page
I. INTRODUCTION.....	1
Significance and Statement of this Work.....	2
Coder Attributes.....	3
G.729 Coder.....	5
G.729 Annex A.....	6
G.729 Annex B.....	7
Organization of this Work.....	8
II. LITERATURE SURVEY.....	10
III. DELAY FOR FIXED RATE CODER.....	15
Coding Delay.....	16
Propagation Delay.....	17
Queuing Delay at the PBX's and Routers.....	17
Receiver Buffer Delay.....	20
Total Delay.....	21
Numerical Results.....	22
IV. DELAY FOR VARIABLE CODER.....	26
Coding Delay.....	26
Propagation Delay.....	26
Queuing Delay.....	27
Receiver Buffer Delay.....	28
Total Delay.....	28
Numerical Results.....	29
Conclusions.....	31
V. ANALYSIS OF VOICE PACKET SIZE AND INTER-ARRIVAL TIME.....	33
PDF of Number of Packets in a Talk spurt.....	34
Voice Packet Inter-Arrival Time Distribution.....	34
Conditional PDF of Voice Packet Size, Given a Talk spurt.....	36
Voice Frames Per Packet in a Fixed Rate Coder.....	41
Conditional PDF of Non-Voice Frames in a Packet, Given a Silence Interval.....	41

PDF of Number of Voice Frames in a Packet.....	42
VI. ANALYSIS OF QUEUE SIZE FOR A SINGLE VOICE SOURCES.....	46
PDF of Queue Size.....	47
Numerical Results.....	52
VII. ANALYSIS OF QUEUE SIZE FOR MULTIPLE VOICE SOURCES.....	61
Method I.....	61
Method II.....	62
Method III.....	63
Numerical Results.....	65
VIII. SUMMARY AND CONCLUSIONS.....	68
REFERENCES.....	72

LIST OF FIGURES

Figure	Page
III-1 Network under study.....	16
III-2 Number of Calls/Mbps vs. Trunk Load.....	23
III-3 Delay vs. Number of Routers for 60% Load.....	24
IV-1 Number of Calls/Mbps vs. Trunk Load.....	30
IV-2 Delay vs. Number of Routers for 60% Load.....	30
IV-3 Delay vs. Number of Routers for 50% Load.....	31
V-1 PDF of number of packets in a talk spurt.....	35
V-2 PDF of inter-arrival time.....	35
V-3 Number of voice frames vs. time for a 2 frame packet.....	36
V-4 Number of voice frames vs. time for a 4 frame packet.....	38
V-5 Conditional Probability Density Function (PDF) of voice packet size for 4 frames/packet given a talk spurt.....	40
V-6 Conditional Probability Density Function (PDF) of pause packet size for a maximum of 4 frames/packet.....	41
V-7 PDF of the packet size for 4 frames/packet for a fixed rate coder.....	44
VI-1 Single voice source model.....	46
VI-2 PDF of number of packets in a talk spurt, 4 frames in a packet.....	47
VI-3 PDF of bits increase in queue during a talk spurt.....	50
VI-4 PDF of queue size (for 1 frame/packet, output line speed of 30 kbps, trunk load of 41%) from this work (top) and OPNET simulation (bottom).....	53
VI-5 PDF of queue size (for 1 frame/packet, output line speed of 25 kbps, trunk load of 49%) from this work (top) and OPNET simulation (bottom).....	54

VI-6	PDF of queue size (for 1 frame/packet, output line speed of 20 kbps, trunk load of 61%) from this work (top) and OPNET simulation (bottom).....	55
VI-7	PDF of queue size (for 1 frame/packet, output line speed of 15 kbps, trunk load of 81%) from this work (top) and OPNET simulation (bottom).....	56
VI-8	PDF of queue size (for 4 frames/packet, output line speed of 8.8 kbps and line load of 61%) from this work (top) and OPNET simulation (bottom).....	57
VI-9	PDF of queue size (for 4 frames/packet, output line speed of 6.6 kbps and line load of 81%) from this work (top) and OPNET simulation (bottom).....	58
VII-1	Multiple voice source model.....	61
VII-2	PDF of queue size using Method II for 1 frame/packet, output line speed of 60 kbps and 3 voice sources.....	66

I. INTRODUCTION

In recent years, rapid advances in different technologies, such as microelectronics, signal processing, switching and transmission, have made possible the deployment of packetized voice over high-speed networks. Although voice traffic has traditionally been handled by circuit switched networks, transmission of voice in packet form offers two key advantages.

Firstly, it may utilize the bursty nature of speech signals to provide a less expensive technique for voice transmission. Variable Bit Rate (VBR) coders have been developed to take advantage of this characteristic of speech. Packets might only be transmitted when a user is speaking. In the absence of speech, during pauses in the conversation, demands on network resources are reduced or completely eliminated. In a circuit switched network such as the Plain Old Telephone System (POTS), this is not the case. Until one of the parties hangs up, network resources are reserved even though no voice traffic may be emanating from one of the sources. Secondly, voice traffic can be integrated with a network designed to move data traffic, thereby increasing the network utilization. This is becoming increasingly important today, given that the volume of data traffic transmitted now surpasses that of voice. The old model of developing a network geared to transmit voice, and integrating data onto it, is becoming less-and-less applicable. Packet networks may also easily take advantage of high compression encoding schemes, which would be prohibitively expensive to implement on POTS, as well as real time and best effort multiplexing.

Significance and Statement of this Work

With the widespread use of the Internet, Voice over the Internet (VoIP), which is based on ITU-T H.323 [1], IETF protocols TCP/IP and RTP [2-4], and ITU-T's G series codecs [5-10], appears to have great promise in the telecommunication systems of future. Until now, little or no work has been carried out for a generalized model of the queue distribution over a packet switched transmission system for ITU-T G.729 series voice coders, prime candidates for carrier VoIP systems. In this study, we develop a mathematical model to examine the queue distribution associated with the real time transportation of Voice over IP using the most recently introduced G.729 series voice coders. This model is easily extendable to other types of coders. We base our work on the statistical model of speech activity developed at Bell Labs by Brady [11]. We perform a worst-case calls supported analysis for a fixed rate coder, i.e. determine the load and the packet size such that the voice packet is delivered end-to-end within 150 ms. This analysis is then extended to include variable rate coders. Finally, we develop a model for predicting the queue distribution for a single voice source and extended it to account for multiple voice sources.

The major issue in VoIP is to maintain the Quality of Service (QoS) required for voice connections. End-to-end delay influences the QoS experienced by end users. It is considered one of the biggest obstacles to VoIP quality. In addition to contributing to the research base in the area of delay distribution associated with the real time transportation of voice over IP using G.729 series coder, the present work is of considerable interest to Williams Communication. It will help the service provider to accommodate the maximum number of voice connections for a given load and number of voice frames in

the packet. This tool will aid in the optimization of their network utilization while maintaining the desired QoS for end users.

Coder Attributes

The speech coder attributes are the most important elements in selecting a coder. Some attributes are more important in one application than another that has a different focus. The specific attributes to be discussed here include bit rate, delay, quality, and complexity. The bit rates that have been in utilization range from 2.4 kb/s to 64 kb/s. The lower rates, 2.4 kb/s and 4.8 kb/s, have generally been used for secure telephony. The digital cellular speech coders cover the midrange of bit rates, where values vary from 6.7 kb/s to 13 kb/s for the Global System for Mobile Communications. Speech coders in this range are also being examined with interest by VoIP providers and manufacturers. The higher range of bit rates, 16 to 64 kb/s, were standardized by the earlier ITU protocols and are aimed at digital telephony applications. Most of the coders are fixed rate coders, that is, they operate at the same fixed rate with no regard to the input.

A second important aspect of voice is the delay. The delay can have a significant impact on the type of application for a coder. Consider, for example, a coder for a real-time conversation. It has been found that in a conversation, if there is a one-way delay of more than 300 ms then it becomes essentially a half-duplex rather than an ordinary conversation [12]. On the other hand, if we compare this situation to a one-way non-interactive voice transmission, then any additional delay in storing and playing it back is less significant as the user in general will not notice it, provided the additional delay is constant. The ITU coders that have the lowest delay are G.711 Pulse Code Modulation (PCM) and G.726 Adaptive Differential PCM (ADPCM). These coders also have the

highest bit rates. In order to digitize voice, speech is typically divided into blocks or frames and then encoded one frame at a time. The frame sizes vary. The first generation coders had frame lengths of 20 ms. Other components of the delay include look ahead, processing delay and transmission delay. The algorithm used for a speech coder, determines the frame size and look ahead (algorithmic delay). The processing delay and the transmission delays are determined by the system. Let us take the example of a cellular system. The digital cellular system has one coder on a single digital signal processing chip. Suppose that the system uses a 20 ms frame size coder with a 5 ms look-ahead. The processing delay for the cellular system will be 20 ms and the total codec delay will be 45 ms, 25 ms to collect the voice samples, followed by 20 ms to process the samples prior to the following frame being presented.

In secure telephony, quality is equivalent with intelligibility. This is the most important requirement for secure telephony. Early speech coders operated on a sample-by-sample basis. The quality was directly related to their Signal to Noise Ratio (SNR) in quantizing the signal samples. At low bit rates, speech was coded based on a speech production model. This model was not equipped to handle anything other than voice. The result was that when the input speech contained background noise, the performance of the speech coder dropped significantly. A considerable amount of research has been done on techniques to make low bit rate speech coders more robust to extraneous noise and signals.

Two additional attributes of speech coders that are important are specification and schedule. The specification of a speech coder varies with the intention of the standards body. In the secure telephony standard, only the bit stream is specified. In

implementation, we may use any coder compatible with the bit stream. In the bit exact specification, the algorithm is fully specified along with the arithmetic precision at each step. The schedule of the standards body, such as ITU, determines whether the coders go through extensive testing or not. If the schedule is shorter, the coders are likely to be derived from previous coders and the testing is very limited. G.723.1 and G.729 Annex A are examples of coders that had short schedules. On the other hand, coders with long schedules are G.721, G.728 and G.729.

G.729 Coder

As was mentioned previously, the G.729 voice coder is considered to be a prime candidate for use in carrier based VoIP telephony systems. Key characteristics of this protocol are discussed here.

ITU-T Recommendation G.729 gives the details of an algorithm based on conjugate structure algebraic code-excited linear-prediction (CS-ACELP) for the coding of speech signals at 8 kb/s. The idea here was to come up with a toll quality 8 kb/s speech coder for wireless applications. The basic requirement also included low delay, low complexity and high quality while operating at 8 kb/s over a noisy channel. The bit rate of 8 kb/s does not include channel coding. One of the trade-offs was between delay and complexity. The result largely met the quality objectives while still delivering a speech coder with low enough complexity. After considerable debate, the frame size was set at 10 ms. The look-ahead delay is 5 ms and the total one-way coding delay is 25 ms (10 ms to capture speech in a frame, 5 ms to capture a portion of the following frame, and finally 10 ms to encode this information before the next frame must be encoded). The initial implementation of the G.729 coder had a complexity of about 20 MIPS and required 3 k

words of RAM. The initial implementation also did not meet some of the performance specifications for loss of frame, also known as frame erasure, and noisy input speech. As the coder was created for wireless channels, it needed to exhibit robustness for both random bit errors and frame erasures. A complete set of test results is provided in [13].

G.729 Annex A

The G.729 Annex A was established for use in digital simultaneous voice and data applications. The complexity was set at 10 MIPS so that the modem algorithm and speech coding algorithm could be implemented on the same processor. Due to its interoperability with G.729, this coder can be used instead of G.729 when complexity reduction is needed in terminal equipment. Some of the other multi-media applications of G.729 include multiparty conferencing, collaborative computing, remote presentations, telemedicine, automated teller machines with voice support, credit card verification and interactive games [14]. Some of the potential applications include Internet telephony and Internet voice mail. The relatively low complexity and delay features make it an attractive choice for such applications compared to G.723.1, which has at least twice the complexity and three times the delay. The low complexity feature of G.729A is important in Internet applications since the algorithm is likely to be run by the host processor in a window based environment in which the processor will be performing other tasks simultaneously. The general description of the G.729A is similar to that of G.729 [12]. It uses the same conjugate structure code-excited linear predictor coding (CS-CELP) concept. The coder operates on speech frames of 10 ms and a sample rate of 8000 samples/s. It has a 5 ms of look-ahead. The speech signal is analyzed after every 10 ms to extract the parameters of the CELP model which are then encoded and transmitted. At

the decoder, these parameters are used to retrieve the excitation and synthesis filter parameters. There is also a G.729 Annex B. It describes a voice activity detector and comfort noise generator to be used for silence compression with either G.729 or G.729 Annex A.

G.729 Annex B

G.729 Annex B defines a low-bit-rate silence compression scheme. It has been designed to work with G.729 and its low complexity Annex A. It is important for digital simultaneous voice and data applications to have further reduction in the bit rate by using silence compression techniques. Silence detection and comfort noise injection results in dual-mode speech techniques, one for speech and the other for silence. A signal classifier, called the Voice Activity Detector (VAD), determines whether the input speech signal is active or inactive. The speech coder operates during active voice speech and a different coding technique, which uses fewer bits, is used during the inactive voice signal.

A voice activity decision, the output of the VAD, is either 1 or 0 indicating an active or inactive signal. This decision is used as a switch between the active or inactive voice encoders. A voice bit stream is generated for each frame when the voice coder is on. A Silence Insertion Description (SID), or nothing, is sent during the inactive periods. G.729 Annex B uses a VAD algorithm along with discontinuous transmission (DTX), SID, and a comfort noise generator (CNG). These algorithms operate under a variety of levels and characteristics of speech and noise. There is a bit-rate saving and no degradation in perceived quality of speech. The coded speech and silence can be transmitted at an average rate of 4 kb/s or less without degradation in overall signal quality as no additional delay is introduced by these algorithms [15].

G.729, G.729A and G.729B have the advantage of lower delay compared to G.723.1. Since the inherent delay of G.723.1 is relatively large, any additional delay will affect the quality of the conversation, or possibly reduce the number of calls the system is capable of supporting given some end-to-end voice delivery delay constraint. G.729 Annex A has the advantage when it comes to complexity. It requires only 10 MIPS and 2000 words of RAM. G.729 Annex B, a voice activity and comfort noise generator procedure, is best suited for bit-rate-sensitive applications. The G.729 series of coders are a complete speech coding package suitable for a wide range of applications in wireless, wireline, satellite communication networks, and Internet and multi-media terminals.

Organization of this Work

The remainder of this work is organized as outlined here. Chapter II provides an outline of related work carried out by other researchers. Some key contributions that have direct correlation to this work are discussed in detail. In Chapter III, we consider a fixed rate coder and carry out a worse case analysis for a 150 ms end-to-end delay. Quantities of interest, such as the number of calls supported and number of voice frames per packet for 150 ms delay, are computed. Chapter IV deals with similar computations but considers a variable rate coder, G.729 Annex B. Here we make use of self-similar traffic queuing theory.

In Chapter V, we map Brady's speech activity model to the G.729B speech coder. We derive the probability density function (PDF) of voice packet size for a variable rate and a fixed rate coder. In Chapter VI, we obtain the PDF of queue size for a single voice source. The results obtained are compared with those of a simulation software (OPNET). The work for a single voice source is extended to account for multiple voice sources in

Chapter VII. The bits in queue when the area under the queue size PDF approaches 99% are compared with results from OPNET. Finally, the important contributions of this work, and areas of further research are summarized in Chapter VIII.

II. LITERATURE SURVEY

In this Chapter, we provide an overview of related work carried out by other researchers in the area of Voice over IP. Wherever possible, brief descriptions of the problems solved and the techniques employed are provided.

The initial research in the transport of non-fixed rate speech over the telephone system was carried out at Bell Labs. In the early 1960's, engineers there proposed a Time Assignment Speech Interpolation (TASI) method for the telephone system that practically doubled the message capacity of existing long submarine cables [16]-[17]. With TASI, many calls share the same facilities, each requiring an available channel only when speech is transmitted. Dubnowski and Crochiere [18] formulated the problem of variable rate coding of speech as a multistate coder coupled with a time buffer. The authors also suggested methods for implementing a variable rate coder based on a dynamic buffering approach. In 1997, Babich [19] and his co-workers presented preliminary results of multiplexing gain that can be achieved by transmission of variable rate speech on a slotted and framed radio channel. Nanda *et. al.* [20] proposed a packet voice protocol for cellular systems.

In the early 1990's, a considerable amount of research was focussed on voice over ATM networks. This work included a self-similar traffic model suitable for analysis of a ATM queue [21], study of the impact of cell delay variation on speech quality [22] and simulation of end-to-end delay distribution [23]. In 1994, Kim and Un [24] analyzed the performance of a statistical multiplexer for bursty voice traffic in an ATM network. The data/voice bursty traffic was modeled by a Markov-modulated Poisson process (MMPP).

A number of other researchers proposed models for simulation of communication links with statistically multiplexing bursty voice [25], video [26]-[29], data [30] or multiple classes of bursty traffic in ATM [31]-[32]. Forgie and Nemeth [33] proposed a Packetized Virtual Circuit (PVC) scheme for integrating voice and data traffic in a communication system. This scheme was used to predict voice delay characteristics for a PVC. In [34], a hybrid technique is presented for estimating input queuing delays that utilizes an analytically derived effective service time distribution at an input queue. The use of variable rate speech coding scheme in ATM environment has been investigated in [35]-[37]. The impact of multimode VBR speech coding on QoS provided by an ATM network is dealt with in [37]. Chandra and Reibman [38] addressed the problems of traffic modeling, estimation of packet delay and statistical multiplexing gain for ATM networks. New strategies and algorithms for transmission and multiplexing of delay-sensitive traffic in packet network were presented in [39]-[42]. In [43], an analytical expression to calculate buffer overflow probability, in which sources have Pareto distributed ON periods and exponentially distributed OFF periods, is given. The authors in [44] present a formula for calculating the decay rate in the queuing system $G/D/1$ which has potential applications in cell based systems such as ATM.

A number of researchers have focussed on the modeling of packetized voice using various models, such as a Markov Chain. Weinstein and Forgie [45] presented an excellent review of speech communication in packet networks. In [46], the authors investigate the real-time efficiency of IP telephony using the IntServ model. A generating function approach to analyze discrete queues with arrival and service processes characterized by Markov Chain is presented in [47]. In [48]-[55] the following topics are

discussed: analytical expression for packet loss probabilities [48]; end-to-end voice call performance which drops the less significant bits in voice packets during periods of congestion [49]; traffic characteristics of packetized voice and its delay performance in a statistical multiplexer [50]; a method for adaptively controlling the flow of voice traffic in an integrated packet network [51]; cut-out fraction (which is the fraction of speech lost due to busy circuits) of a TASI system [52]; superposition model to analyze a finite buffer statistical multiplexer with multiple arbitrary on/off input sources [53]; mathematical model for obtaining performance characteristics of a finite buffer packet voice multiplexer and fractional packet loss using an embedded Markov Chain [54]; and a uniform arrival and service model to analyze a packet speech multiplexer [55].

Several researchers have obtained measurements for delay and packet loss on the Internet, while others have primarily targeted the area of queuing analysis of data traffic. Borella and Brewster [56] analyzed 12 traces of round-trip Internet data packet delay. The traces exhibit Hurst parameter estimates greater than 0.5, indicating long-range dependence. Several other researchers [57]-[59], have also discussed the implications of long-range dependence in network traffic. A number of experimental studies [60]-[67] have focussed on end-to-end Internet data traffic dynamics by tracing packets exchanged between a large campus network, a state-wide educational network and a large Internet service provider. In [68], Tang *et. al.* provides a Pareto queuing model for internet data flow traffic. The worst case network delay performance of a self-clocked fair queuing scheme is studied in [69]. In [70], multiscale queuing analysis that provides a simple closed form approximation to tail queue probability, valid for any given buffer size, is provided for data traffic. Girish and Hu [71] proposed a Markov modulated arrival

processes to approximate the correlated arrival processes in the Internet. Casetti and Meo [72] evaluated the queuing delay and packet loss of TCP flows using an M/D/1/B queue model.

In [73], a comparison of delay and jitter performance of VoIP traffic generated by different standard voice codec algorithms is provided. A multi-rate speech coder which includes concealment and correction techniques against loss of packets in VoIP transmission is presented in [74]. Miyata *et. al.* [75] proposed a new concept called “Loss Window Size” that extends the definition for packet loss and delay to groups of packets. The authors also present information on packet loss and packet delay measured on the Internet. A tutorial on transmission engineering (with emphasis on VoIP) covering the topics of speech coding, packetization and transportation is provided in [76]. Kostas *et. al.* [77] examined possible architectures for VoIP and discussed measured Internet delay and loss characteristics. In [78], Mishra and Saran address the problem of designing capacity management and routing mechanisms to support telephony over an IP network. Hoshi *et. al* [79] proposed an RTP voice stream multiplexing method for IP gateways. In [80], the author focuses on understanding the cause of delay within analog modems, with the objective of developing recommendations to minimize delay for VoIP applications. An overview of some technical problems associated with the provisioning of interoperable services between IP telephony and the PSTN is provided in [81].

Most recently, a group of researchers have provided results on average delay on various types of networks. A closed form expression for delay in packetized voice transport using an M/G/1 queue model is given in [82]. In [83]-[90], the authors have provided average delay computations for voice transport over satellite internet access

networks [86] and VoIP networks [83], [85], [88], and QoS for packetized voice over Universal Mobile Telecommunication System air interface [84], [87]. In [91] and [92] the authors analyze the multiplexing of a large/small number of independent and homogeneous on-off traffic sources into a single packet switched buffer. The authors employ an excess rate technique [44], which has been used successfully in the analysis of a number of queuing systems, to the analysis of VoIP using G/D/1 input queue. A text by Minoli [93] provides a review of VoIP. It contains a review of IP technologies, discussion of voice characteristics, overview of vocoder-based compression methods used in IP and protocols for delivering of voice in IP environments.

The speech models developed by Paul Brady at Bell Labs may be considered the basis for much of the work in this field. In [94], Brady provides numerical results such as talkspurt and pause distribution for various detector thresholds. He also provides speech parameters such as mean talkspurt and mean pause. An extensive set of data on the analysis of on-off speech patterns in 16 experimental telephone conversations is provided in [95].

In reviewing the previous work, we find that little or no work has been carried out for a generalized model of the queue distribution over a packet switched transmission system for voice coders. The present work is expected to make a significant contribution by developing a mathematical model to examine the queue distribution associated with the real time transportation of Voice over IP.

III. DELAY FOR FIXED RATE CODER

In this Chapter, we determine the end-to-end delay for the network under study. The various delay components that compromise this end-to-end delay are outlined. The coder used in this work is ITU-T Recommendation G.729 Annex A, a fixed rate coder, which has a voice frame size of 10 ms and a bit-rate of 8 kbps. According to ITU-T Recommendation G.114 [96], the upper bound on end-to-end delay should be 150 ms for most user applications. Numerical results to be presented here include the number of calls supported as a function of trunk load, and delay vs. number of routers for various loads. These results allow us to determine the number of calls that can be supported for a given load and a fixed end-to-end delay of 150 ms. In addition, we can also determine the maximum number of frames that can be put in a packet for a fixed load and delay. Delay bounds for VoIP using ETSI E-model have been discussed in [83] and [85]. This model is not being used in this work.

The test case network under study is shown in Figure III-1. This network shows typical devices involved with transmitting a Voice over IP telephone call. The network has phones connected to PBX's. The PBX's compress and packetize the voice, and ship the packets to a corporate edge router over a network that is likely carrying a mix of voice and data traffic. The corporate edge router segregates the voice traffic and forwards these packets to a carrier's VoIP network. Other similar local networks are connected only to the first backbone router of the carrier's network. The voice traffic from all the local networks is transmitted to the appropriate destination edge router and PBX, where the packets are uncompressed and a reconstructed voice signal is fed to the phone.

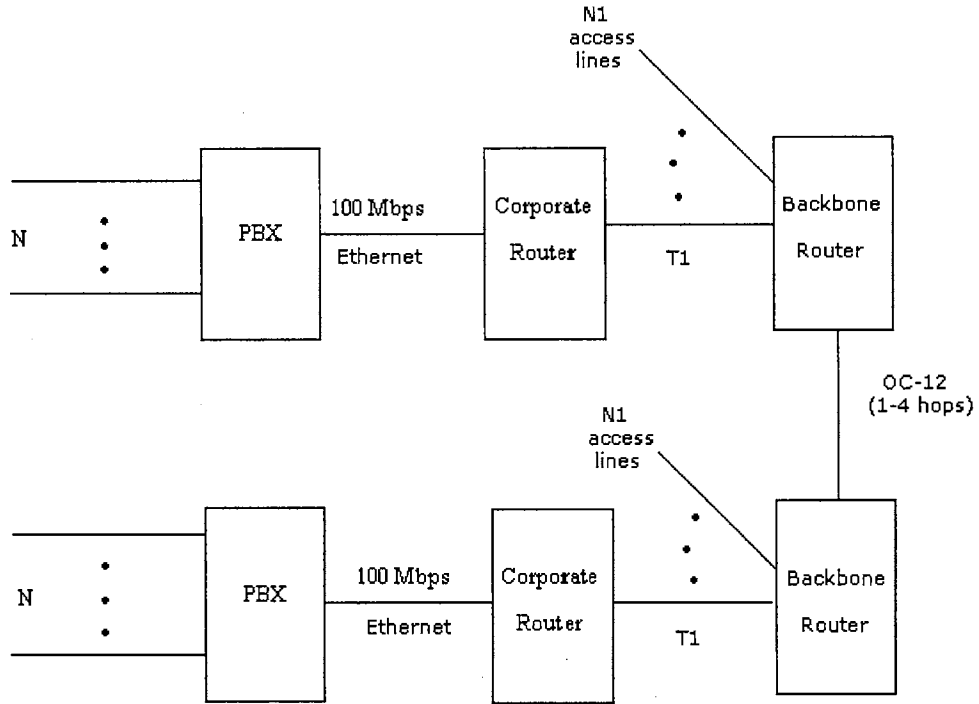


Figure III-1: Network under study

Coding Delay

A voice coder processes a frame of speech at a time and introduces two types of delays- algorithmic delay and processing delay. Before the speech can be analyzed, the entire input frame must be available. This results in delay equal to the frame size. The codec also analyzes samples, which are in the next input subframes. In the G.729A codec, the subframes are equal to 5 ms in duration [8]. This results in a look-ahead delay. Algorithmic delay comprises these two delays. Therefore, the algorithmic delay is a function of frame size and sub-frame size. The processing delay is upper bounded by the frame size [83], [93]. The coding delay, t_c , is equal to the sum of the algorithmic delay, t_a , and the processing delay, t_p . The algorithmic delay is the sum of frame size, t_f , and

subframe size, t_{sf} . The processing delay is the product of the number of frames in a packet, N_f , and the frame size, t_f . Thus, the coding delay can be further written as

$$t_c = (N_f + 1) \times t_f + t_{sf} \quad (\text{III-1})$$

Propagation Delay

The propagation delay, t_{pr} , is given by

$$t_{pr} = t_{pbx} + t_{access} + t_{trunk} \quad (\text{III-2})$$

where t_{pbx} is the propagation delay between PBX's and the corporate routers, t_{access} is the access line propagation delays, and t_{trunk} is the propagation delay of the trunks. t_{pbx} is generally small and can be ignored. Therefore, the propagation delay depends only on t_{access} , and t_{trunk} . It can be expressed as

$$t_{pr} = [(R-1) \times d_r + 2d_a] \times (1/6c) \quad (\text{III-3})$$

R is the number of backbone routers, d_r is the distance between the backbone routers, d_a is the distance covered by access lines, and c is the speed of light (3×10^8 m/s).

Queuing Delay at the PBX's and Routers

We consider a system that moves nothing but voice packets. This is how many of the carriers are currently ensuring QoS for their VoIP traffic. The VoIP traffic is segregated from the data. Note that the analysis procedure used in this section can be easily extended to account for a mix of data and prioritized voice traffic. The queuing delay at PBX is assumed to be small and is ignored.

For the worst case scenario, we assume that the access lines are fully loaded and each source generates a voice packet at the same time. The interval between packet transmission, t_{int} , is given by

$$t_{int} = (N_f)(t_f) \quad (III-4)$$

The packet size in bits, P_s , is dependent on the coder bit rate (S_c), frame size (t_f), and the length of the IP header (47 bytes) and is given by

$$P_s = (N_f)(S_c)(t_f) + 376 \quad (III-5)$$

The number of users supportable by the access line, N , can now be calculated as

$$N = (S_l)(t_{int})/P_s \quad (III-6)$$

where S_l is the available bandwidth of the access line in bits/sec. Therefore, the worst case queuing delay for the first corporate router, t_{cr1} , is given by

$$t_{cr1} = (N)(T_s) \quad (III-7)$$

where T_s is the time it takes to service a packet. $T_s = P_s/S_l$. Substituting for T_s and equation (III-6) in equation (III-7) we get

$$\begin{aligned} t_{cr1} &= (S_l)(N_f)(t_f)/P_s \times (P_s/S_l) \\ &= (N_f)(t_f) \end{aligned} \quad (III-8)$$

From equation (III-4) we find that, t_{cr1} , can be further written as

$$t_{cr1} = t_{int} \quad (III-9)$$

To prevent queues which increase without bounds, the worst case delay through the first corporate router must not exceed the time that a voice source takes to generate a packet, else the previous round of packets will not be cleared by the time next round is generated. The number of access lines supportable by the carrier backbone, N1, can be computed by

$$N1 = (S_t)(\rho)/(S_l) \quad (\text{III-10})$$

where S_t is the available trunk bandwidth, $S_t = (774/810) \times (\text{Trunk Speed})$. The factor $(774/810)$ accounts for a 36 byte overhead in a 810 byte SONET frame. ρ is the load on the trunk. The time to service a packet at the backbone router, T_{sb} , is given by

$$T_{sb} = (P_s)/S_t \quad (\text{III-11})$$

Now, we can calculate the worst case delay at the first backbone router, t_{br1} , as

$$t_{br1} = (N1)(T_{sb}) \quad (\text{III-12})$$

Substituting for N1, and T_{sb} from equations (III-6), (III-10) and (III-11), we get

$$\begin{aligned} t_{br1} &= [(S_t)(\rho)/S_l] \times [P_s/S_t] \\ &= (\rho)(P_s/S_l) \end{aligned} \quad (\text{III-13})$$

In case, the number of routers is greater than 2 ($R > 2$), then we have the next router delay, t_{brn} , which can be written as

$$t_{brn} = (R-2) \times (T_{sb}) \quad (\text{III-14})$$

The last router delay, t_{blr} , is given by

$$t_{blr} = T_s \quad (III-15)$$

In the above calculations, we are assuming that there is no queue at the last (corporate) router. This is a reasonable assumption if voice traffic is designed to have high priority and the LAN speed is considerably greater than the access line speed.

The total queuing delay, t_q , is

$$t_q = t_{cr1} + t_{br1} + t_{brn} + t_{blr} \quad (III-16)$$

Receiver Buffer Delay

The receiver buffer is required at the receiving end to suppress delay variations exhibited by an incoming stream of voice packets. To insure that the receiver buffer does not empty, traffic should be stored for a time equal to the maximum end-to-end delay variability prior to playing anything back. In a system mixing voice traffic with data traffic, or mixing together variable rate voice traffic, the variable delays or jitter will vary from one received packet to another. A system multiplexing together nothing but fixed rate voice sources which are using identical coders may, depending upon the accuracy of the source's clocks, settle into a steady state which exhibit no jitter. Even in this case jitter will occur as individual calls terminate or are established and a voice conversation's packets shift positions in the stream. If a packet is the last one in the queue, the queuing time taken by the packet to reach the other end is equal to the maximum end-to-end queuing time, t_{max} .

Using equations (III-7), (III-12), (III-14 through 16), t_{\max} can be written as

$$t_{\max} = (N)T_s + (N1)T_{sb} + (R-2)T_{sb} + T_s \quad (\text{III-17})$$

If a packet is always first in the queue, the time taken by the packet is equal to the minimum end-to-end queuing time, t_{\min} . The queuing delay at the each router will be equal to the service time. Hence, t_{\min} is given by

$$t_{\min} = 2T_s + T_{sb} + (R-2)T_{sb} \quad (\text{III-18})$$

Realizing that the end-to-end propagation delays are identical for these two cases, the receiver buffer size should be set equal to the difference between (III-18) and (III-17) to insure the buffer does not empty. The receiver buffer delay, t_{rec} , is therefore given by

$$t_{\text{rec}} = (N-1)T_s + (N1-1)T_{sb} \quad (\text{III-19})$$

Total Delay

The total delay, T_d , is the sum of the coding delay, propagation delay, queuing delay and receiver buffer delay. T_d is given by

$$T_d = t_c + t_{pr} + t_q + t_{\text{rec}} \quad (\text{III-20})$$

Substituting for all the delay components using equations (III-1), (III-3), (III-9), (III-12), (III-14)-(III-16) and (III-19), we arrive at

$$T_d = (N_f + 1) t_f + t_{sf} + [(R-1) d_r + 2d_a]/(.6c) + t_{\text{int}} + (N1)T_{sb} + T_{sb} (R-2) \\ + T_s + (N-1)T_s + (N1-1)T_{sb}$$

Substituting for t_{int} from equation (III-4) and T_{sb} from equation (III-11) in the above equation and noting that $(N)T_s = N_f(t_f)$ (ref. equations III-7 and III-8) and $(N_1)T_{\text{sb}} = \rho(P_s/S_i)$ (ref. equations III-12 and III-13), we get

$$T_d = (3N_f + 1)t_f + t_{\text{sf}} + [(R-1)d_r + 2d_a]/(.6c) + 2\rho P_s/S_i + (R-3)(P_s/S_i) \quad (\text{III-21})$$

If $T_d \leq 150$ ms, we calculate the percentage of voice in a packet, V_p , using the following equation

$$V_p = (S_c)(t_f)(N_f)/P_s \quad (\text{III-22})$$

The useable bandwidth, B_u , is given by

$$B_u = V_p \times \rho \quad (\text{III-23})$$

The number of calls supported by the trunk, N_c , can therefore be calculated by

$$N_c = B_u/S_c$$

noting that S_c represents the voice bandwidth required per call. Substituting (III-22) and (III-23) in the above equation, we get

$$\begin{aligned} N_c &= (\rho/S_c)[S_c \times t_f \times N_f]/P_s \\ &= (N_f)(t_f)(\rho)/(P_s) \end{aligned} \quad (\text{III-24})$$

Numerical Results

In this section, we present some sample computations of the total delay. We first define the numerical values of the network variables for this example. G.729A has 10 ms

frame size and 5 ms of look-ahead delay. The local network is 100 Mbps Ethernet. The access lines are T1's with bandwidth of 1.54 Mbps. The access line distance is 10 km. The trunks are OC-12 with bandwidth of 622 Mbps. The maximum number of backbone routers is 5. The distance between the routers is 1000 km. The values of parameters are $t_f = 10$ ms, $t_{sf} = 5$ ms, $d_r = 1000$ km, $d_a = 10$ km, $S_c = 8$ kbps, $S_l = 1.536$ Mbps, $S_t = (774/810)(622 \text{ Mbps})$, $S_{in} = 100$ Mbps. Figure III-2 shows the number of calls/ Mbps as a function of trunk load (ρ). The trunk load is varied from 1% to 90% and $R=5$. In the figure, we show the results for $40\% \leq \rho \leq 90\%$. We varied the number of voice frames per packet, as long as the total delay (T_d) remained below 150 ms for all loads.

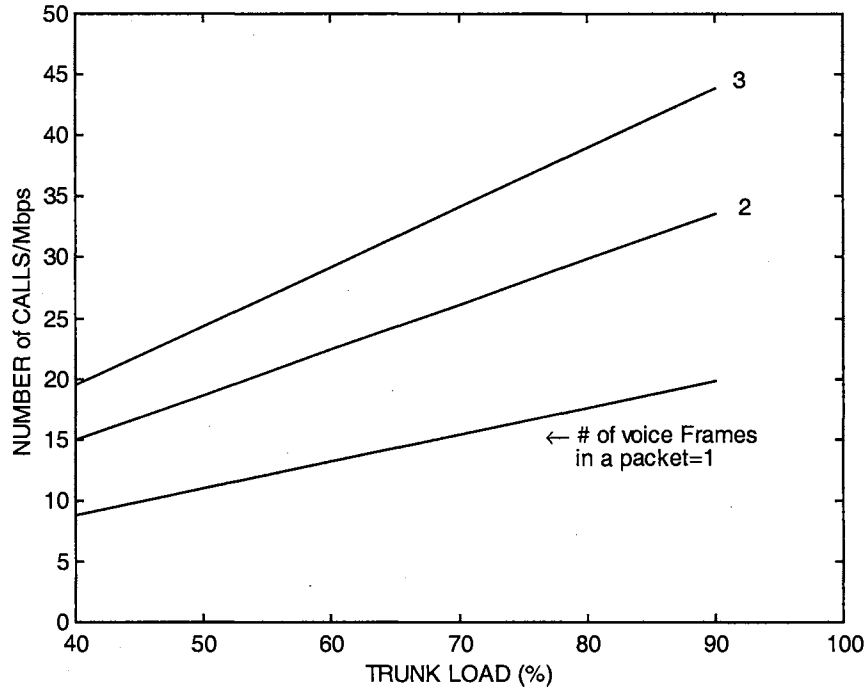


Figure III-2: Number of Calls/Mbps vs. Trunk Load

It can be observed from the Figure III-2 that under the constraints of this simulation, the maximum number of frames allowable in a packet is 3 for $40\% \leq \rho \leq 90\%$

and 4 backbone hops. The maximum number of calls/Mbps is 43.831. We can put 4 frames per packet, but the number of allowable hops drops to 2. The maximum number of calls for this case at $\rho = 90\%$ is 54.598 per Mbps.

In Figure III-3, the total delay is shown as a function of number of backbone routers for different frame sizes. The trunk load is fixed at 60%.

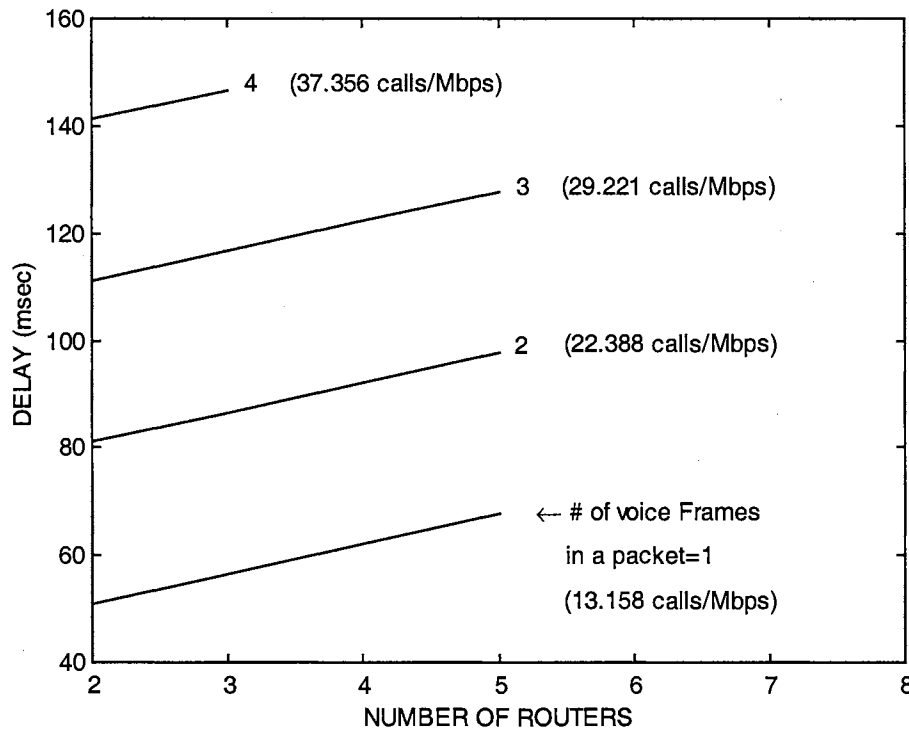


Figure III-3: Delay vs. Number of Routers for 60% Load

In the computation of results of Figure III-3, we observed that for 4 frames in a packet the delay exceeds 150 msec after 2 hops. For $40\% < \rho < 90\%$, we get similar graphs as shown in Figure III-3 except that the delay value differs by 0.3 ms to 0.4 ms.

In this Chapter, we analyzed the end-to-end delay for a test case network using a G.729 Annex A fixed rate coder. A mathematical model was developed to determine

absolute end-to-end delivery time. Numerical results obtained in this Chapter show that for a fixed rate coder and trunk loads ranging up to 90%, we cannot put more than 3 voice frames per packet for 4 hops and expect to meet the end-to-end delay criterion.

IV. DELAY FOR VARIABLE RATE CODER

In this Chapter, we estimate the end-to-end delay using a variable rate coder instead of fixed rate coder, for the network considered in Figure III-1 of Chapter 3. The coder used is ITU-T Recommendation G.729 Annex B, with silence suppression. It too has the voice frame size of 10 ms. The bit rate is 8 kbps when there is voice activity. A Silence Insertion Description (SID) frame or nothing is sent during the silence period. Thus, it has an average bit rate of 4 kbps or less [15]. As mentioned previously, variable rate coders offer potential advantages to VoIP carriers, the most notable being the increased utilization of network bandwidth. The characteristics of variable rate voice coders require that a statistical approach be used to estimate performance, thus complicating the analysis. The upper bound on end-to-end delay is again set at 150 msec. Self-similar traffic queuing theory is used to estimate queuing delays, with a Hurst parameter, $H = 0.85$ [98], a value in the range of that found for actual Internet traffic. Numerical results in this Chapter include the number of calls supported as a function of backbone trunk load, and delay vs. number of routers for various loads.

Coding Delay

The coding delay, t_c , is a function of algorithmic delay and processing delay. It is defined in equation (III-1)

Propagation Delay

The propagation delay, t_{pr} , is defined in the Chapter III and is given by equation (III-3).

Queuing Delay

The queuing delay is the sum of delays at all the routers in the network. The packet size, P_s , is given by equation (III-5). The time it takes to service a packet on the corporate router attached to the access line, T_s , is given by

$$T_s = P_s / S_l \quad (IV-1)$$

where S_l is the available bandwidth of the access line in bits/sec. The average delay for the first corporate router [99], t_{cr1} , is given by

$$t_{cr1} = (T_s \times \rho_a^{(2H-1)/(2-2H)}) / (1-\rho_a)^{H/(1-H)} \quad (IV-2)$$

where ρ_a is the load on access lines. The time to service a packet at the backbone router is given by

$$T_{sb} = P_s / S_t \quad (IV-3)$$

where S_t is the available trunk bandwidth, $S_t = (774/810) \times (\text{Trunk Speed})$.

The average delay at the first backbone router can now be calculated as

$$t_{br1} = (T_{sb} \times \rho_t^{(2H-1)/(2-2H)}) / (1-\rho_t)^{H/(1-H)} \quad (IV-4)$$

where ρ_t is the load on the trunks. In case, the number of routers is greater than 2, the next router delay, t_{brn} , can be calculated from equation (III-14). The last router delay, t_{blr} , is defined by equation (III-15).

The total queuing delay, t_q , is

$$t_q = t_{cr1} + t_{br1} + t_{brn} + t_{blr}$$

Substituting equations (IV-2), (IV-4), (III-14) and (III-15) in the above equation we get

$$t_q = (T_s \times \rho_a^{(2H-1)/(2-2H)})/(1-\rho_a)^{H/(1-H)} + (T_{sb} \times \rho_t^{(2H-1)/(2-2H)})/(1-\rho_t)^{H/(1-H)} + (R-2)T_{sb} + T_s \quad (IV-5)$$

Receiver Buffer Delay

It was noted in the last chapter that the receiver buffer is required at the receiving end to suppress delay variations exhibited by an incoming stream of voice packets. Chapter 3 deals with hard guaranteed bounds whereas this Chapter deals with statistical bounds on end-to-end delay. One problem faced with setting the proper size of the receiver buffer here is that the PDF of the delay distribution for the voice packets is unknown, in fact, one goal of this research is to determine it. For the purpose of this Chapter and to give the reader an idea as to how statistical multiplexing affects the analysis, the receiver buffer delay, t_{rec} , is somewhat arbitrarily set at twice the queuing delay, t_q , given in equation (IV-5). Thus,

$$t_{rec} = 2[(T_s \times \rho_a^{(2H-1)/(2-2H)})/(1-\rho_a)^{H/(1-H)} + (T_{sb} \times \rho_t^{(2H-1)/(2-2H)})/(1-\rho_t)^{H/(1-H)} + (R-2)T_{sb} + T_s] \quad (IV-6)$$

Total Delay

The total delay, T_d , is the sum of the coding delay, propagation delay, queuing delay and receiver buffer delay. Substituting equations (III-1), (III-3), (IV-5) and (IV-6) for all delay components in equation (III-20), we arrive at

$$T_d = (N_f + 1) t_f + t_{sf} + [(R-1) d_r + 2d_a]/(.6c) + 3[(T_s \times \rho_a^{(2H-1)/(2-2H)})/(1-\rho_a)^{H/(1-H)} \\ + (T_{sb} \times \rho_t^{(2H-1)/(2-2H)})/(1-\rho_t)^{H/(1-H)} + (R-2)T_{sb} + T_s]$$

Substituting equations (IV-1) and (IV-3) in the above equation, we get

$$T_d = (N_f + 1) t_f + t_{sf} + [(R-1) d_r + 2d_a]/(.6c) + 3(\rho_a^{(2H-1)/(2-2H)})/(1-\rho_a)^{H/(1-H)} P_s/S_l + \\ 3(\rho_t^{(2H-1)/(2-2H)})/(1-\rho_t)^{H/(1-H)} P_s/S_t + 3(R-2) P_s/S_t + 3P_s/S_l \quad (IV-7)$$

The number of calls supported by the trunk, N_c , can be calculated from equation (III-24).

Numerical Results

Sample computations of the total delay are presented in this section. The numerical values used in computations are same as those used in the results of Chapter 3. The coder bit rate, S_c , is taken equal to 4 kbps. Figure IV-1 shows the number of calls/Mbps as a function of trunk load (ρ_t). The access line load (ρ_a) is set equal to the trunk load (ρ_t). The trunk load is varied from 10% to 90% and the number of routers is set equal to 5. We varied the number of voice frames per packet, as long as the average total delay (T_d) remained below 150 ms.

From Figure IV-1, we observe that we can put 9 frames per packet for an upper limit on the load of 45% to satisfy the end-to-end delay criterion. The number of calls supported for this case is 73.905 per Mbps. We can put up to 3 frames per packet for a load up to 60%. For load equal to 40% or less, we can put 10 frames per packet for 4 allowable hops. In this case the number of calls decreases to 68.027 per Mbps. We can even put 11 frames in a packet for 4 allowable hops as long as the load is 15% or less and the number of calls in this case decreases to 26.274 per Mbps.

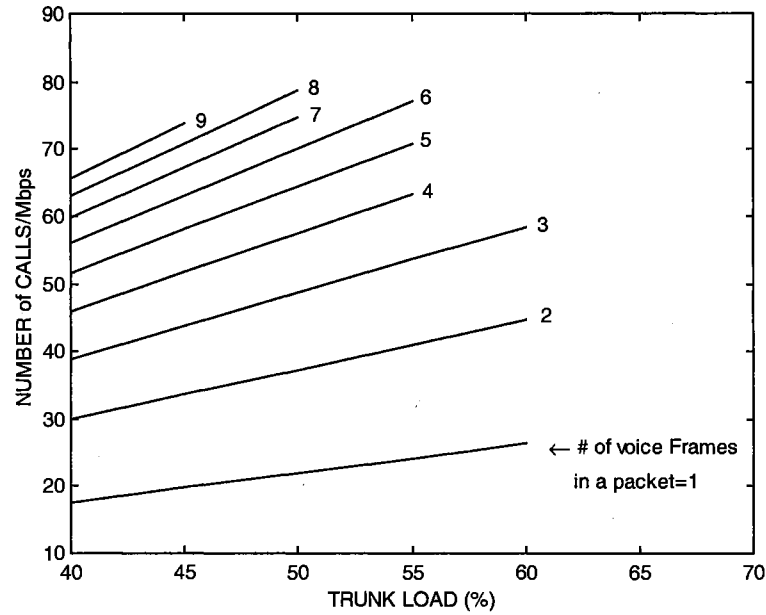


Figure IV-1: Number of Calls/Mbps vs. Trunk Load

In Figure IV-2, total delay is shown as a function of number of routers for different frame sizes. The load is fixed at 60%.

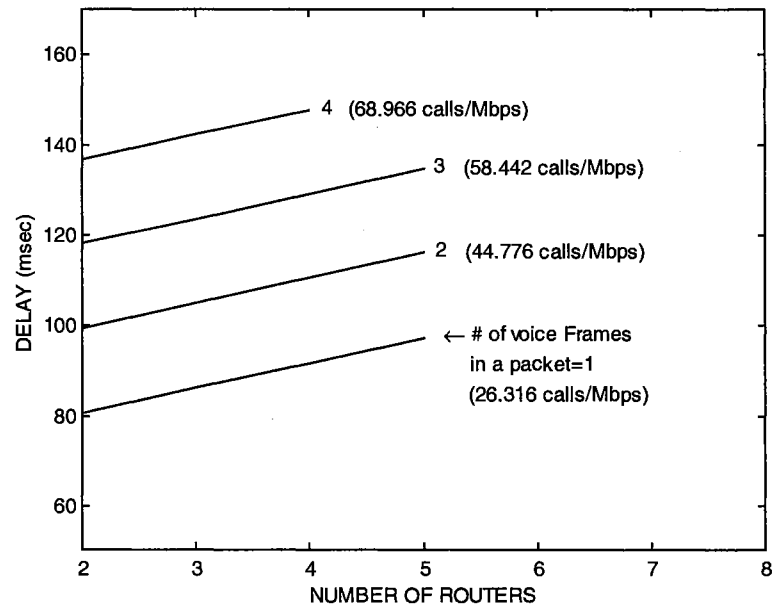


Figure IV-2: Delay vs. Number of Routers for 60% Load

In Figure IV-2, we find that to meet the delay criterion 3 voice frames can be put in the packet for 4 allowable hops. We can put 4 frames per packet, but the number of allowable hops decreases to 3. In Figure IV-3, delay vs. number of routers for 55% load is shown.

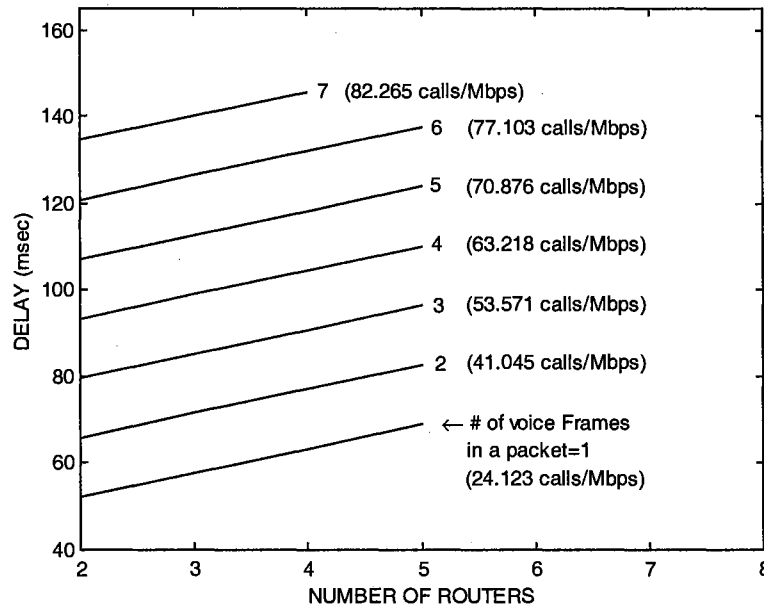


Figure IV-3: Delay vs. Number of Routers for 55% Load

From this Figure, we see that for 7 frames/packet we cannot have more than 3 hops.

Conclusions

At a first glance, the reader might think that there is a tremendous performance improvement for variable rate voice coders, but the numerical results in Chapter 3 for fixed rate coders are based on absolute end-to-end delivery times and the results in Chapter 4 are based on average end-to-end delivery times. If the end-to-end delay PDF is symmetrical, 50% of the packets in the analysis in Chapter 4 will arrive over the target end-to-end delivery time.

The model of Chapter 3 is adequate for designing VoIP networks using a fixed rate coder, as they are based on absolute values of end-to-end delivery time. However, the model of Chapter 4 is not adequate for designing variable rate VoIP networks as it is based on average values. A high quality variable rate coder system may need to deliver 99% of its packets within the 150 ms delivery limit, and right now we have no way to estimate this. Therefore, we need to analyze the queue distribution of real time transportation of variable rate voice over IP for a more accurate design of this type of network. In the next chapter, we begin to address this issue.

V. ANALYSIS OF VOICE PACKET SIZE AND INTER-ARRIVAL TIME

In the previous two Chapters, we focussed our attention on the design of VoIP networks based on absolute end-to-end delivery time for fixed rate VoIP coders, and average end-to-end delivery times for variable rate VoIP coders. It was noted that the latter analysis needs to be refined in order to better model the statistical characteristics of networks carrying variable rate Voice over IP traffic. In reviewing the literature, we note that little or no such related work has been carried out so far. Previous studies have either been focussed on voice over ATM or the modeling of delay distributions using Markov chains, M/M/1, or M/D/1 queuing models. These models do not accurately represent the arrival and service processes of voice traffic over IP. In this Chapter, we develop a generalized model of the distributions of the voice packet size, and also note the voice packet inter-arrival times.

We base our work on the speech activity model developed at Bell Labs by Paul Brady, which alternates talk spurts with silence intervals [11]. We use the cumulative distribution function (CDF) of the talkspurt from Brady's model which is given by

$$F(t) = 1 - e^{-\alpha t} \quad (V-1)$$

where α^{-1} is the mean of the talkspurt and t is the length of the talkspurt in seconds. The CDF of the silence is same as equation (V-1), except that α is replaced by β . β^{-1} is the mean of a silence interval. From the Brady's analysis [95], we take $\alpha^{-1} = 1.125$ seconds and $\beta^{-1} = 1.721$ seconds. The probability density function (PDF) of the talk spurt is then obtained by differentiating equation (V-1) and is given by

$$f(t) = \alpha e^{-\alpha t} \quad (V-2)$$

PDF of Number of Packets in a Talk Spurt

We can define the pdf of the number of voice packets generated given a talk spurt of length t via

$$S_k = P[(k-1)t_p \leq t < kt_p] \quad (V-3)$$

where S_k denotes the probability of k packets and t_p is the packet size in seconds. The packet size, t_p , is given by

$$t_p = n_f \times t_f \quad (V-4)$$

where n_f is number of frames in a packet and t_f is the frame size in seconds.

With the use of equations (V-1) and (V-2), we can now rewrite S_k in equation (V-3) as

$$\begin{aligned} S_k &= \int_{(k-1)t_p}^{kt_p} f(t) dt \\ &= F(kt_p) - F((k-1)t_p) \end{aligned} \quad (V-5)$$

We use equation (V-5) to generate the PDF of the number of voice packets in a talk spurt for a packet size of 4 frames. A portion of this PDF is shown in Figure V-1 for a frame size, $t_f = 10$ ms.

Voice Packet Inter-Arrival Time Distribution

The CDF of the inter-arrival time for voice packets is taken to be [100]

$$F(t) = [(1-\alpha T) + \alpha T(1-e^{-\beta(t-T)})] u(t-T) \quad (V-6)$$

where T is the packet generation time in seconds. The PDF of the inter-arrival time is determined from equation (V-6) and is given by

$$f(t) = (1-\alpha T)\delta(t-T) + \alpha\beta T e^{-\beta(t-T)} u(t-T) \quad (V-7)$$

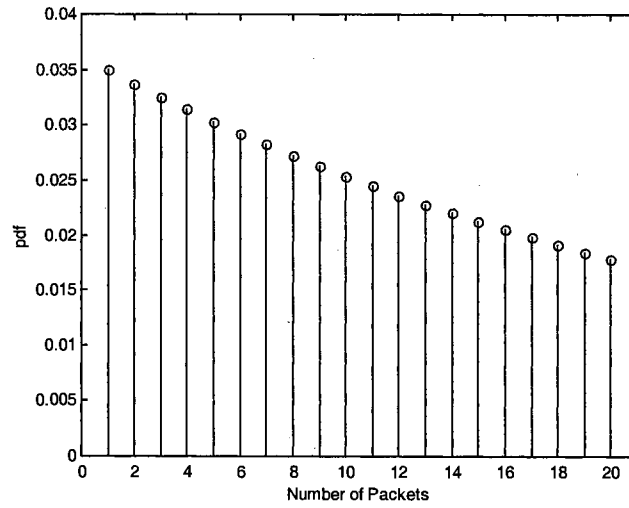


Figure V-1: PDF of number of packets in a talk spurt

Figure V-2 shows the PDF of the interarrival time for $T = 40$ ms (4 frames/packet).

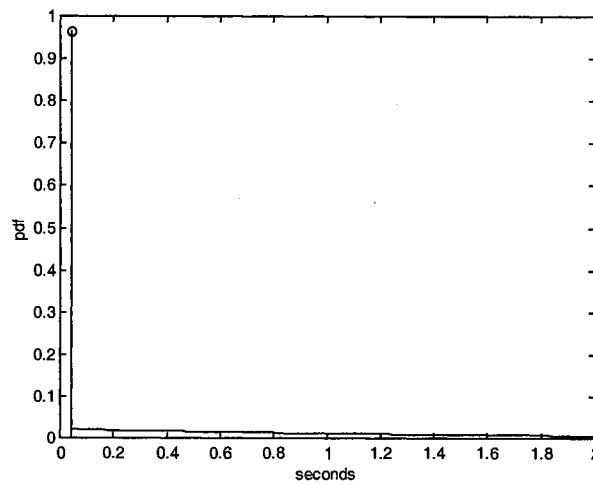


Figure V-2: PDF of inter-arrival time

From this figure, we find that the PDF is dominated by the large spike at $t = T$. The exponent component is seen to be very small and decays slowly. The shape of this PDF can be explained as follows. During a talkspurt, a large number of packets will arrive at T second intervals, hence the large spike at time T . After a talkspurt, the time to the next packet arrival depends on the length of the silence interval and is exponentially distributed.

Conditional PDF of Voice Packet Size, Given a Talk Spurt

To determine the PDF of the voice packet size given a talk spurt, we draw a graph showing the number of frames for a given packet size vs. time in ms.

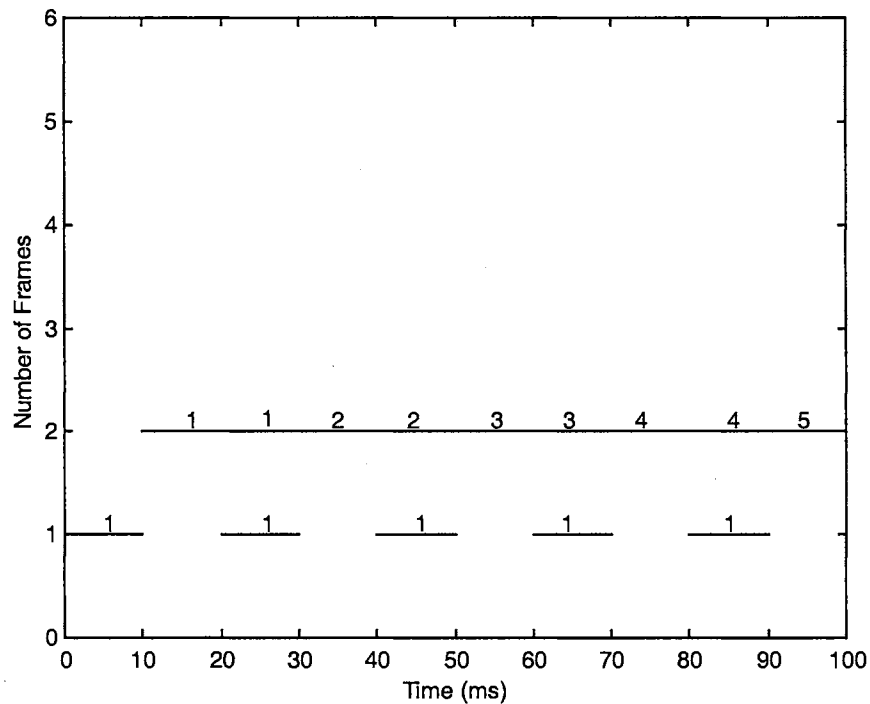


Figure V-3: Number of voice frames vs. time for a 2 frame packet

Figure V-3 shows the two voice frames per packet case. The number on each horizontal line denotes the number of packets. For example, for $0 < t < 10$ ms we get 1 packet with 1 frame of voice, whereas for $40 < t < 50$ ms we get 2 packets containing 2 voice frames/packet, and 1 smaller packet containing 1 voice frame. We define p_m as the probability that the length of the talk spurt is between $(m)t_f$ and $(m-1)t_f$ seconds, let $E_k(n)$ be number of packets with k frames in a packet of maximum size n expected during a talk spurt. From the figure V-3, we can determine this expected value as

$$E_1(2) = p_1 + p_3 + p_5 + p_7 + p_9 + \dots$$

$$= \sum_{i=1}^{\infty} p_{2i-1},$$

Similarly, the expected number of packets in a talk spurt with maximum sized 2 frames per packet is

$$E_2(2) = p_2 + p_3 + 2p_4 + 2p_5 + 3p_6 + 3p_7 + 4p_8 + 4p_9 + \dots$$

$$= \sum_{i=1}^{\infty} \sum_{j=1}^2 (i) p_{2i+j-1},$$

Let $P_k(n)$ be the probability of getting k frames in a packet with the maximum packet size of n frames. The probability of getting 1 frame in a packet with maximum size of 2 frames is

$$P_1(2) = E_1(2) / \left[\sum_{i=1}^2 E_i(2) \right]$$

and the probability of getting 2 frames is

$$P_2(2) = E_2(2) / \left[\sum_{i=1}^2 E_i(2) \right]$$

The above example is extended for a packet size of 4 frames, as shown in Figure V-4.

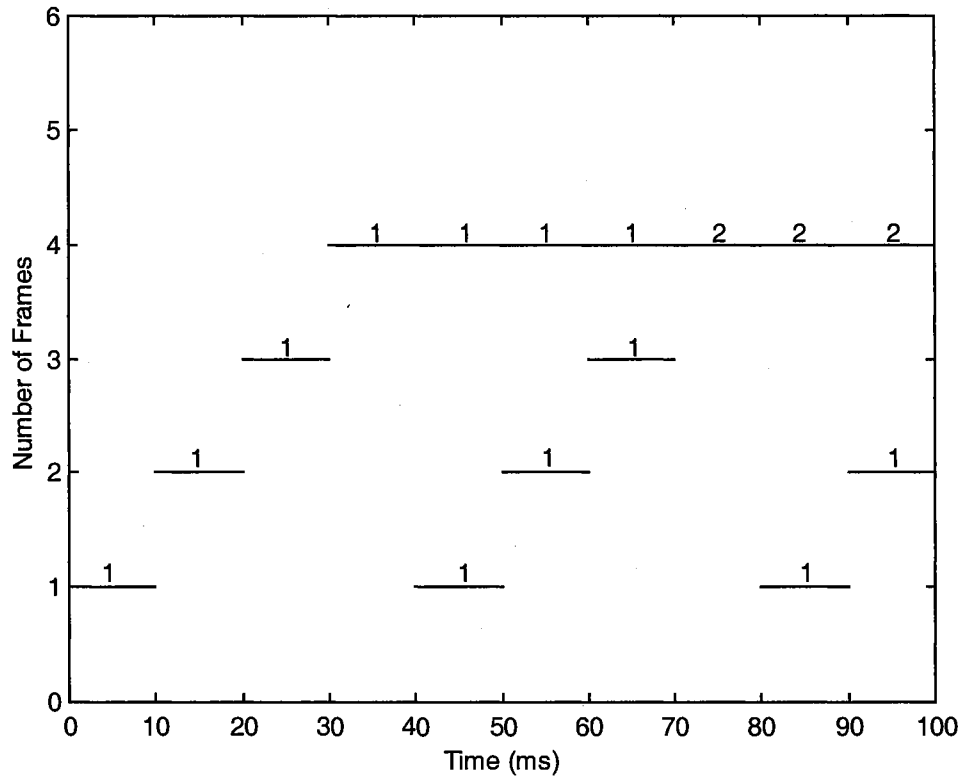


Figure V-4: Number of voice frames vs. time for a 4 frame packet

From this figure, we get the expected value of getting a packet with 1, 2, 3, or 4 frames as:

$$E_1(4) = p_1 + p_5 + p_9 + p_{13} + p_{17} +$$

$$= \sum_{i=1}^{\infty} p_{1+4(i-1)},$$

$$E_2(4) = p_2 + p_6 + p_{10} + p_{14} + p_{18} + \dots$$

$$= \sum_{i=2}^{\infty} p_{2+4(i-2)},$$

$$E_3(4) = p_3 + p_7 + p_{11} + p_{15} + p_{19} + \dots$$

$$= \sum_{i=3}^{\infty} p_{3+4(i-3)},$$

$$E_4(4) = p_4 + p_5 + p_6 + p_7 + 2p_8 + 2p_9 + 2p_{10} + 2p_{11} + 3p_{12} + \\ 3p_{13} + 3p_{14} + 3p_{15} + 4p_{16} + \dots$$

$$= \sum_{i=1}^{\infty} \sum_{j=1}^4 (i)p_{4i+j-1},$$

Now we can calculate probability of getting 1, 2, 3, or 4 frames as

$$P_k(4) = E_k(4) / \left[\sum_{i=1}^4 E_i(4) \right], \quad k=1, 2, 3, 4$$

With the help of these examples, we can generalize the equation for the probability of voice packet size.

$$E_k(n) = \sum_{i=k}^{\infty} p_{k+n(i-k)}, \quad k = 1, 2, \dots, n-1 \quad (V-8)$$

$$E_n(n) = \sum_{i=1}^{\infty} \sum_{j=1}^n (i)p_{ni+j-1} \quad (V-9)$$

$$p_m = \int_{(m-1)t_f}^{mt_f} f(t) dt \quad (V-10)$$

$f(t)$ is the PDF of talk spurt given in equation (V-2), n is the maximum number of frames in packet, k is the number of frames, and t_f is frame size in seconds.

The probability of getting k frames in a packet of size n is given by

$$P_k(n) = E_k(n) / \left[\sum_{i=1}^n E_i(n) \right], \quad k = 1, 2, \dots, n \quad (V-11)$$

Using equations (V-8)-(V-11), we can now compute the conditional PDF of voice packet size given a talk spurt.

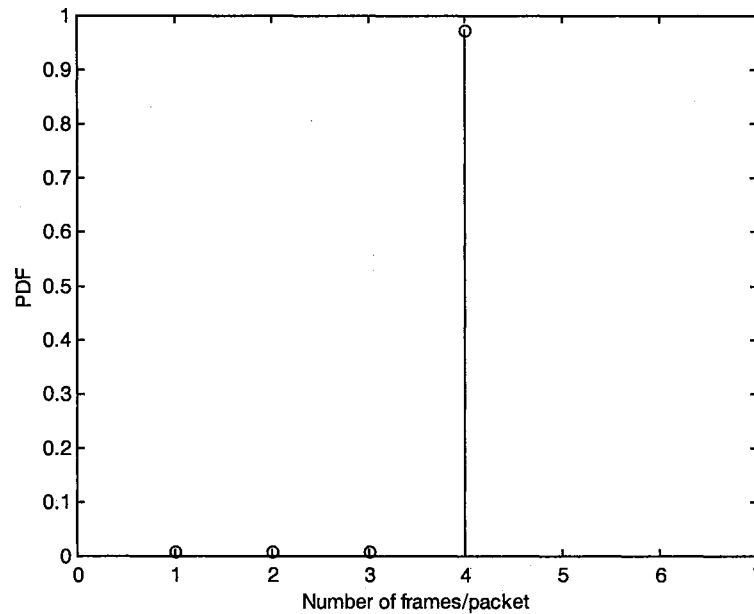


Figure V-5: Conditional Probability Density Function (PDF) of voice packet size for 4 frames/packet given a talk spurt

Figure V-5 shows the result when the maximum number of frames in a packet is four. In this case, the probability of getting 4 frames of voice in a packet is 0.9737,

whereas the probability of getting 1, 2, or 3 frames in a packet is 0.0088, 0.0088 and 0.0087, respectively.

Voice Frames Per Packet in a Fixed Rate Coder

The Figure V-5 gives us the PDF of voice packet for a variable rate coder. In this section we derive the voice packet size PDF for a fixed rate coder. In case of the fixed rate coder, packets of a fixed size containing non-voice background noise are sent during silence intervals.

Conditional PDF of Non-Voice Frames in a Packet, Given a Silence Interval

The conditional PDF of the silence packet size can be defined by equations (V-8)-(V-11). The $f(t)$ in equation (V-10) is replaced by pause PDF given by equation (V-2) with α replaced by β .

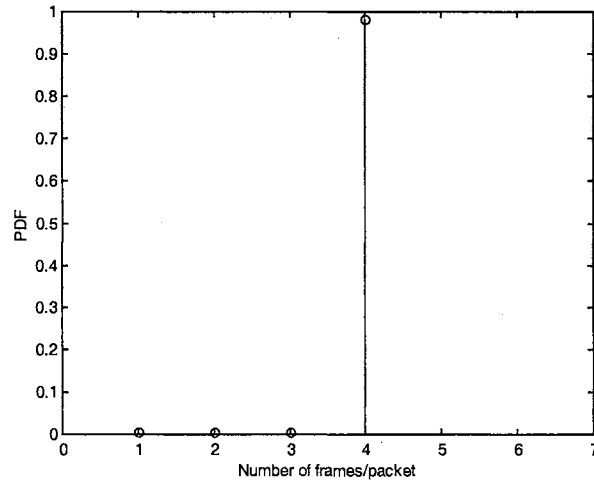


Figure V-6: Conditional Probability Density Function (PDF) of pause packet size for a maximum of 4 frames/packet

Figure V-6 shows the conditional PDF of pause packet size of 4 frames. The frame size is 10 ms. The figure shows that the probability of getting 4 frames of silence in

a packet is 0.9827, whereas the probability of getting 1, 2, or 3 frames of silence in a packet is 0.0058, 0.0058, and 0.0057, respectively.

PDF of Number of Voice Frames in a Packet

It should be pointed out that the graph of the number of pause frames as a function of time is similar as the one for voice (as shown in Figure V-4). Let p_i^s be the probability that the length of silence is between $(i)t_f$ and $(i-1)t_f$. The first step to determine the unconditional PDF of the number of voice frames in a packet is to flip the graph of the number of pause frames as a function of time. I.E. in Figure V-6 four silence frames in a packet equals zero voice frames, three silence frames means there was one voice frame, etc. We can then obtain the following expected values of getting 0 through 4 frames of voice.

$$E_0^s(4) = p_4^s + p_5^s + p_6^s + p_7^s + 2p_8^s + 2p_9^s + 2p_{10}^s + 2p_{11}^s + 3p_{12}^s + 3p_{13}^s + 3p_{14}^s + 3p_{15}^s + 4p_{16}^s + \dots$$

$$= \sum_{i=1}^{\infty} \sum_{j=1}^4 (i)p_{4i+j-1}^s,$$

$$E_1^s(4) = p_3^s + p_7^s + p_{11}^s + p_{15}^s + p_{19}^s + \dots$$

$$= \sum_{i=1}^{\infty} p_{3+4(i-1)}^s,$$

$$E_2^s(4) = p_2^s + p_6^s + p_{10}^s + p_{14}^s + p_{18}^s + \dots$$

$$= \sum_{i=2}^{\infty} p_{2+4(i-2)}^s,$$

$$E^s_3(4) = p^s_1 + p^s_5 + p^s_9 + p^s_{13} + p^s_{17} + \dots$$

$$= \sum_{i=3}^{\infty} p^s_{1+4(i-3)},$$

Now, we can calculate the probability of getting 0, 1, 2, 3 frames as

$$P^s_k(4) = E^s_k(4) / \left[\sum_{i=0}^3 E^s_i(4) \right], \quad k = 0, 1, 2, 3$$

We can now write the above equations in generalized form.

$$E^s_k(n) = \sum_{i=k}^{\infty} p^s_{(n-k)+n(i-k)} \quad k = 1, 2, \dots, n-1 \quad (V-12)$$

$$E^s_0(n) = \sum_{i=1}^{\infty} \sum_{j=1}^n (i) p^s_{ni+j-1} \quad (V-13)$$

$$p^s_m = \int_{(m-1)t_f}^{mt_f} s(t) dt \quad (V-14)$$

where $s(t)$ is the PDF of the pause given by equation (V-2) with α replaced by β .

$$P^s_k(n) = E^s_k(n) / \left[\sum_{i=0}^{n-1} E^s_i(n) \right], \quad k = 0, 1, \dots, n-1 \quad (V-15)$$

Since we know that a typical conversation comprises of 40% of actual talking and 60% of silence, we normalize the equation (V-11) by a factor of 0.4 and equation (V-15) by a factor of 0.6. By adding these normalized equations, we get the generalized equations for the packet size PDF given by

$$P_k^p(n) = 0.4 \times (E_k(n) / [\sum_{i=1}^n E_i(n)]) + 0.6 \times (E_k^s(n) / [\sum_{i=0}^{n-1} E_i^s(n)]), \quad k = 1, 2, \dots, n-1 \quad (V-16)$$

$$P_0^p(n) = 0.6 \times (E_0^s(n) / [\sum_{i=0}^{n-1} E_i^s(n)]), \quad (V-17)$$

$$P_n^p(n) = 0.4 \times (E_n(n) / [\sum_{i=1}^n E_i(n)]) \quad (V-18)$$

The equation (V-16) gives the probability of getting a 1, 2, ... n-1 voice frames in packet. Equation (V-17) gives the probability of getting no voice frames in a packet and equation (V-18) gives the probability of getting a packet full of voice frames. Using equations (V-8)-(V-10), (V-12)-(V-14), and (V-16)-(V-18), we obtain the PDF of packet size for 4 frames/packet shown in Figure V-7.

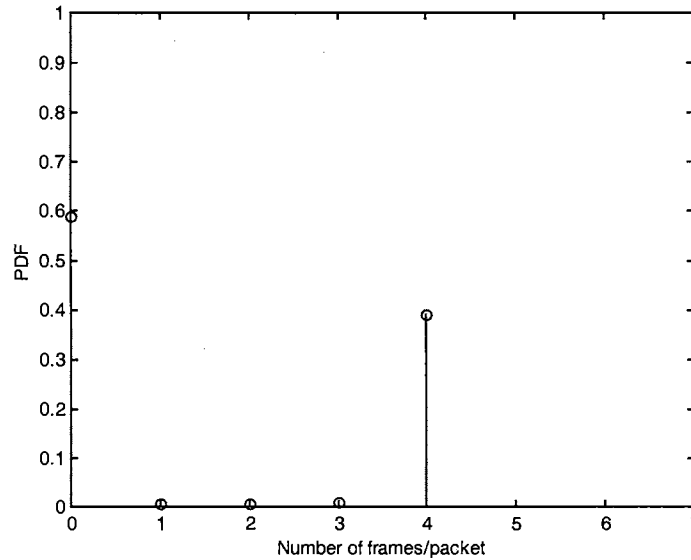


Figure V-7: PDF of the packet size for 4 frames/packet for a fixed rate coder

Figure V-7 shows the probability of getting a packet with 0, 1, 2, 3, or 4 frames of voice in it for a fixed rate coder inserting four voice frames in each packet. In this example, the probability of getting an no voice frames is 0.5896 and the probability of getting 4 frames of voice in a packet is 0.3895. The probability of getting 1, 2, or 3 frames of voice in a packet is 0.0070, 0.0070, and 0.0070, respectively.

VI. ANALYSIS OF QUEUE SIZE FOR A SINGLE VOICE SOURCE

In Chapter V, we used the cumulative distribution function of the talk spurt from Brady's model to determine the probability density function of the talk spurt. The resulting PDF of the talk spurt was used to arrive at the PDF of the packet size. In reviewing the literature, we note that previous work has either focused on voice over ATM or modeling of queue distributions using M/M/1 or M/D/1 models. These models do not accurately represent the arrival and service processes of voice traffic over a packet network. In this Chapter, we compute the PDF of the queue size in a packet switch resulting from a single voice source.

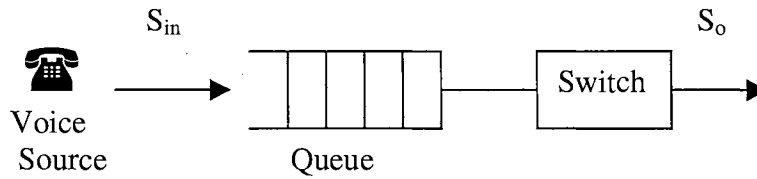


Figure VI-1: Single voice source model

Consider the model shown in Figure VI-1. The switch has an input line speed of S_{in} and output line speed of S_o , and switching occurs in a store-and-forward manner. The voice source is bursty, alternating between talk spurts and silence intervals. During a talk spurt, numerous packets arrive at a fixed inter-arrival interval of T seconds. If the output line speed is less than the average input rate of the voice traffic, the queue will begin to fill. During a silence period, the input line will be idle and the queue will commence emptying.

Using the formulation presented in Chapter V (equations V-1 through V-5), we generate a PDF of the number of packets in a talk spurt. For example, the probability of k packets, denoted by S_k and given by equation (V-5) is calculated and shown in Figure VI-2 for an assignment of 4 frames per packet.

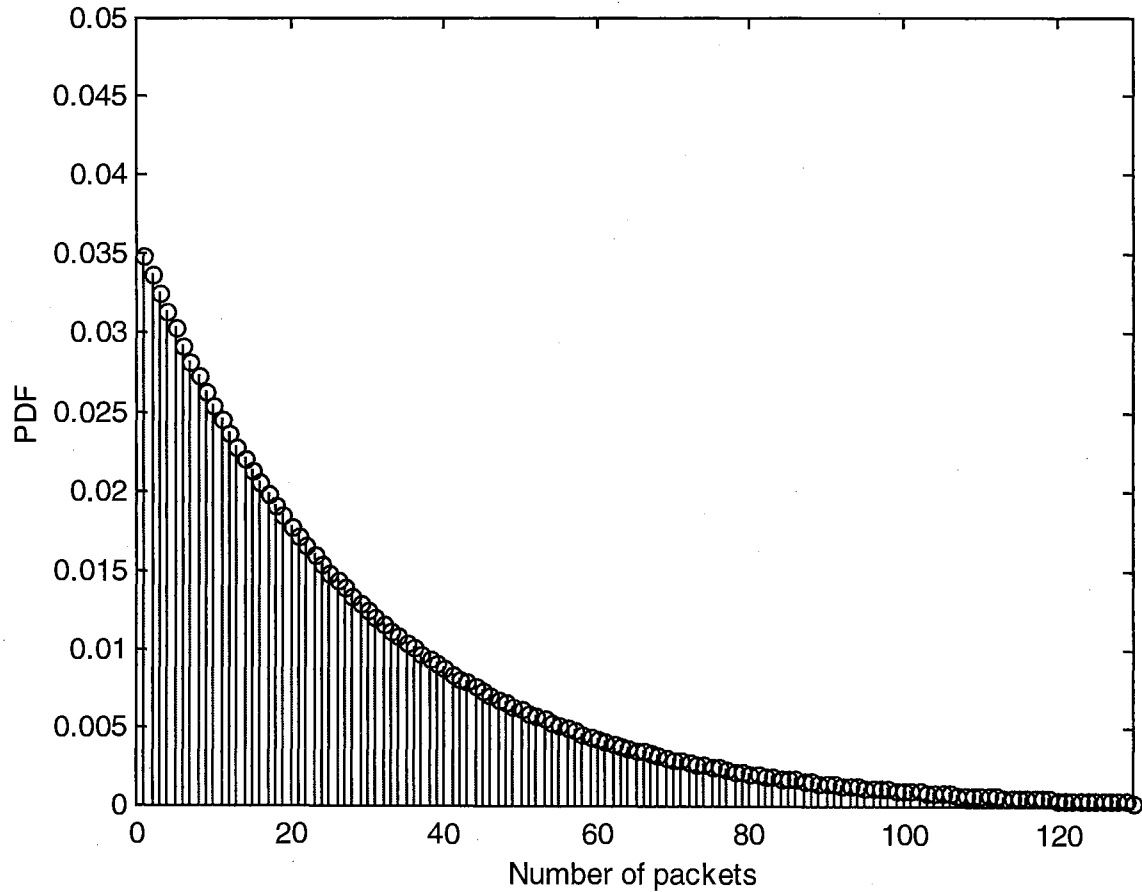


Figure VI-2: PDF of number of packets in a talk spurt, 4 frames in a packet

PDF of Queue Size

The packet stream from a single voice source during a talk spurt is characterized by arrivals at fixed intervals of T ms given by

$$T = t_f \times n_f \quad (\text{VI-1})$$

where t_f is the frame size in ms and n_f is the number of frames in a packet.

We define packet injection time, T_{in} , as

$$T_{in} = P_s / S_{in} \quad (\text{VI-2})$$

where S_{in} is the input line speed and P_s is the packet size in bits and is given by

$$P_s = (t_f \times n_f) S_c + \text{overhead} \quad (\text{VI-3})$$

where S_c is the coder speed. The packet size PDF derived in Chapter 5, shows that the number of voice frames in all packets will be essentially the same. In other words, we get a full packet or nothing. The packet service time, T_{out} is given by

$$T_{out} = P_s / S_o \quad (\text{VI-4})$$

where S_o is the output line speed.

Note that in order for a queue to be formed during a talk spurt, $T_{out} > T - T_{in}$. Assuming that the queue is empty to start with, and noting that in a store-and-forward switch the entire packet is stored before it is forwarded, the increase in bits in the queue during the time period T upon the arrival of the first packet is given by

$$P_1 = P_s - (T - T_{in}) S_o \quad (\text{VI-5})$$

When the second packet arrives, the queue is not empty any more and as a result the cumulative increase in bits in the queue in a time period T due to the arrival of the subsequent packets is given by

$$P_i = P_1 + (i-1)(P_s - TS_o), \quad i = 2, 3, \dots \quad (\text{VI-6})$$

In this case, a preceding packet is exiting the queue simultaneous with the arrival of packet i .

Equations (VI-5) and (VI-6) can be used to determine the PDF of the bit increase in a queue during a talk spurt, if the talk spurt arrives at an empty queue. Technically, if the talk spurt arrives at a partially filled queue, a different PDF based solely on a slightly modified equation (VI-6) is the most accurate form to use. In the derivation of this chapter, the PDF derived from (VI-5) and (VI-6) is used throughout, as experimentation has revealed that these two PDF versions yield negligible difference in the end result.

Using equations (VI-1) through (VI-3), (VI-5) and (VI-6), the PDF of the number of packets in a talk spurt can be converted into a PDF of the bit increase in the queue during a talk spurt. This is carried out by converting the x axis from 'number of packets' to 'bits increase' by replacing the first packet by P_1 obtained from equation (VI-5) and the subsequent packets by P_i , $i=2, 3, \dots$, obtained from equation (VI-6). The mean, λ_d^{-1} , for the discrete PDF of bits increase in queue, is given by

$$\lambda_d^{-1} = \sum_i P_i p_i \quad (\text{VI-7})$$

where p_i is the probability that event P_i occurs. Figure VI-3 shows the discrete PDF of the bits increase in the queue for $S_o = 8.8$ kbps, $T = 40$ ms, $S_{in} = 1.54$ Mbps, $S_c = 8$ kbps, $t_f = 10$ ms, $n_f = 4$, and $P_s = 544$ bits. The mean, λ_d^{-1} , for this example is 5476 bits.

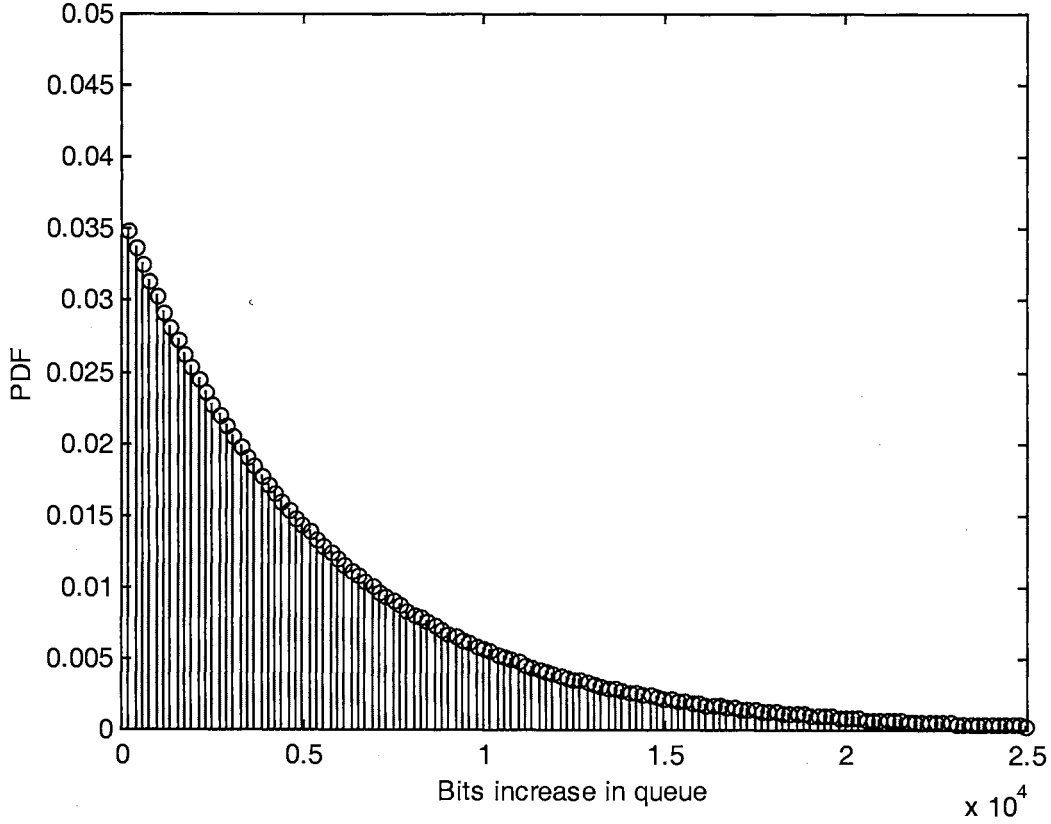


Figure VI-3: PDF of bits increase in queue during a talk spurt

In the similar manner, we can arrive at the discrete PDF of the bit decrease in the queue as follows. We first determine the probabilities S_k , given in equation (V-5) with α^{-1} given in equation (V-2) replaced by β^{-1} , where β^{-1} is the mean of the pause, to arrive at a PDF similar to the one shown in Figure VI-2. Then we scale the x-axis by $-TS_o$ to arrive at the discrete PDF of bits decrease in the queue. The factor, TS_o , represents the bit decrease between each n_{ftf} intervals. Then we calculate the mean, γ_d^{-1} , of the discrete PDF of bit decrease in the queue.

For the purpose of further computations which will use Fast Fourier Transforms with equal spaced sample points not necessarily spaced as the PDF of the queue bit increase during a talk spurt and the queue bit decrease during a silence interval, we model

the discrete PDF of the bit increase in queue during a talk spurt by a continuous exponential function given by

$$f_b(x) = \lambda e^{-\lambda(x-P_1)}, \quad x \geq P_1 \quad (\text{VI-8})$$

where λ^{-1} is the mean increase in bits, x is the number of bits and P_1 is given by equation (VI-5). We set the mean of $f_b(x)$, $\lambda^{-1} + P_1 = \lambda_d^{-1}$, where λ_d^{-1} is the mean of the discrete PDF of bit increase in the queue.

Similarly, we approximate the PDF of bit decrease in the queue that occurs during a silence interval by another continuous exponential function

$$f_d(x) = \gamma e^{\gamma(x+TS_o)}, \quad x \leq -TS_o \quad (\text{VI-9})$$

The mean of $f_d(x)$, $\gamma^{-1} + TS_o$, is set equal to the mean of the discrete PDF of bit decrease in the queue, γ_d^{-1} .

Now, we outline the procedure to determine the queue PDF

Step 1: Compute $f_1(x) = f_b(x) \otimes f_d(x)$

where $f_1(x)$ is the convolution of $f_b(x)$ and $f_d(x)$ given by equations (VI-8) and (VI-9), respectively.

Step 2: All points $x < 0$ in $f_1(x)$ are mapped to $x = 0$. This gives $f_2(x)$.

Step 3: Compute $f_3(x) = f_2(x) \otimes f_b(x)$

and

$$f_4(x) = f_3(x) \otimes f_d(x)$$

$$\text{Set } f_1(x) = f_4(x)$$

Repeat steps 2 and 3 until steady state is reached. $f_4(x)$ gives us the statistical PDF at the *end* of the silence interval.

Step 4: Compute $f_5(x) = f_4(x) \otimes f_b(x)$

$f_5(x)$ gives us the statistical PDF at the *end* of the talk spurt interval.

Step 5: Compute $f_6(x)$ which is the average of PDF's $f_4(x)$ and $f_5(x)$ obtained in Step 3 and Step 4, respectively.

The rationale for this process is explained as follows. Define a random variable Y to be the number of bits in the queue, X_b as the increase in bits during a talk spurt, and X_d as the decrease in bits during a silence interval. Noting that talk spurts and silence intervals alternate, and assuming that these random variables are statistically independent, we can write the number of bits in the queue after one on-off cycle as $Y = X_b + X_d$. The PDF of the Y can be found by convolving $f_b(x)$ with $f_d(x)$. Noting that the queue size Y cannot be negative, any portion of the PDF of Y that is less than zero after this convolution is mapped to zero. To examine the effects of multiple on-off cycles, this process can be repeated until the result reaches a steady state. The steady state result of $f_4(x)$ gives the statistical PDF of queue size when ending with a silence whereas $f_5(x)$ gives the statistical PDF of queue when ending with the talk spurt. The average of the two PDF's, $f_4(x)$ and $f_5(x)$, can be used to approximate the time average queue PDF, which is what we are interested in for real-world queuing analysis. The above procedure can be carried out efficiently using the Fast Fourier Transform.

Numerical Results

Here, we present some results of the queue size for a single voice source. Figures VI-4 through VI-7 show the queue size PDF for a 1 frame/packet experiment. The

parameters used for these cases are $T = 10$ ms, $n_f = 1$, $t_f = 10$ ms, $P_s = 304$ bits, and $S_{in} = 1.54$ Mbps. The figures show the result based on the theory presented in this chapter (top) and OPNET simulation (bottom). In comparing the results of this work with the OPNET simulation, we find good agreement between simulation and theory. The general shape of the PDF's matches quite well.

We used $S_o = 30$ kbps and line load = 41% for Figure VI-4. The theory shows a spike of 0.944 at $x = 0$. This spike at the origin means that the queue is empty 94.4% of the time. In comparison, OPNET simulation shows that the queue is empty 93% of the time.

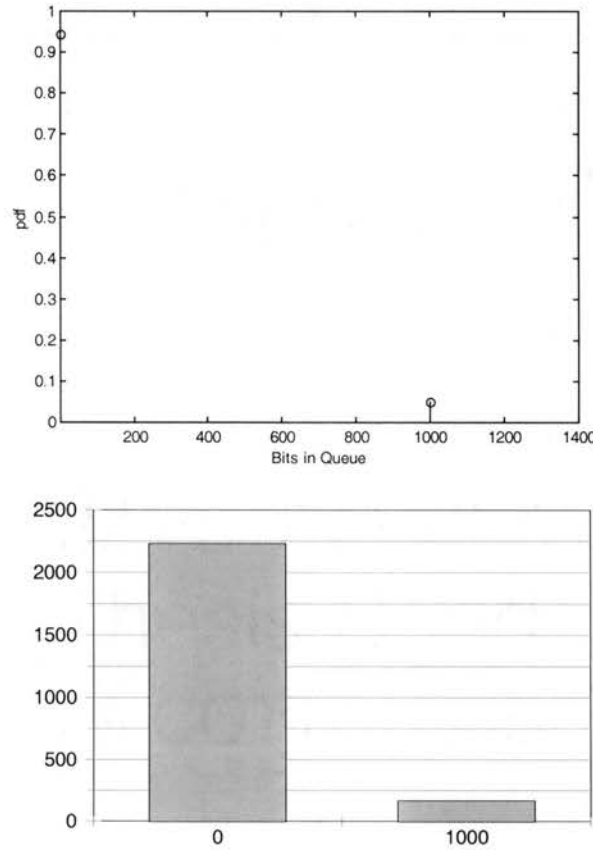


Figure VI-4: PDF of queue size (for 1 frame/packet, output line speed of 30 kbps, trunk load of 41%) from this work (top) and OPNET simulation (bottom)

In Figure VI-5, $S_o = 25$ kbps, and a line load = 49%. In this case, the queue is empty 51.1% and 55% of the time for this work and OPNET, respectively.

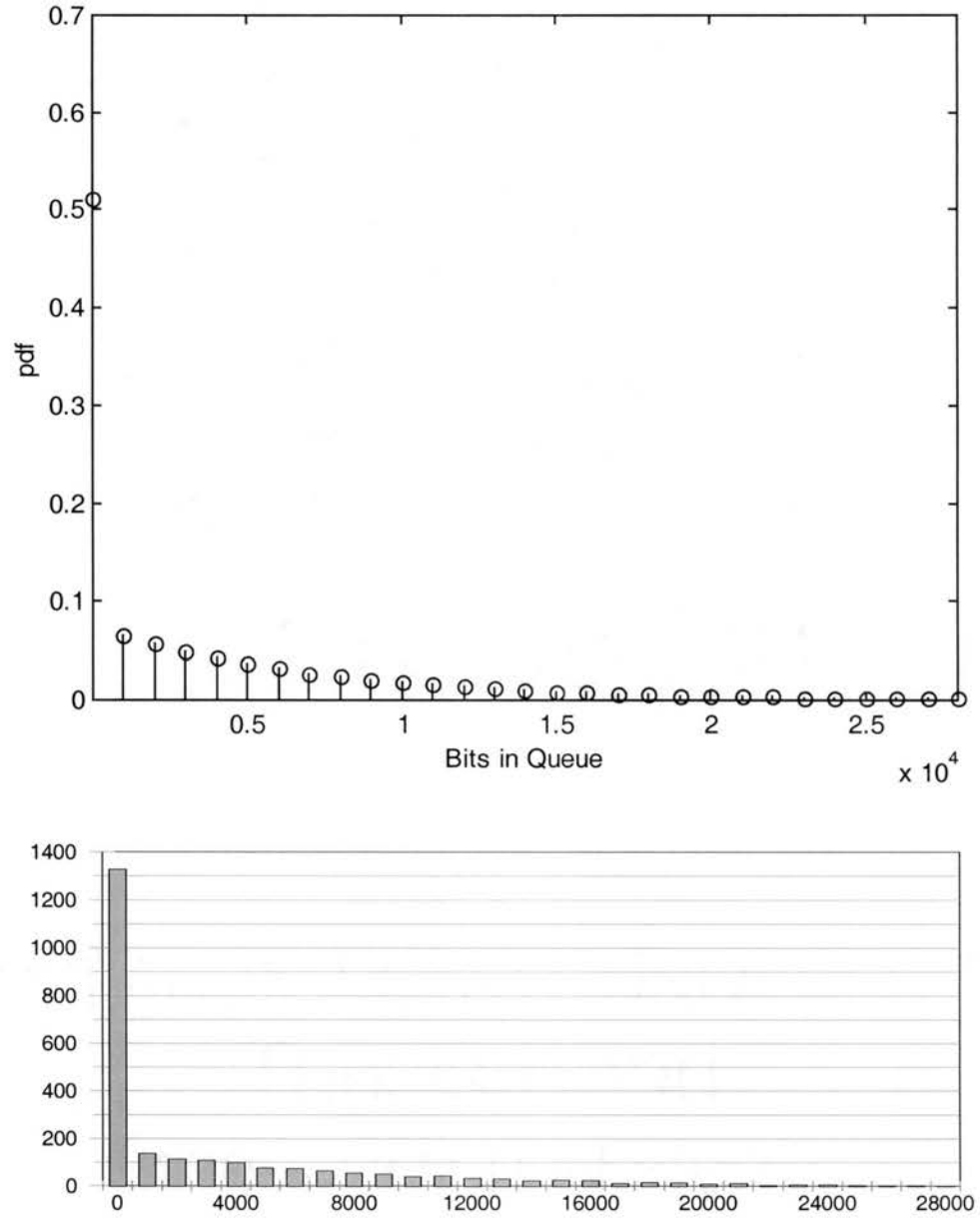


Figure VI-5: PDF of queue size (for 1 frame/packet, output line speed of 25 kbps, trunk load of 49%) from this work (top) and OPNET simulation (bottom)

Figure VI-6 shows the result for $S_o = 20$ kbps and line load = 61%. The queue in this case is empty 38% of the time for this work and 42% for OPNET.

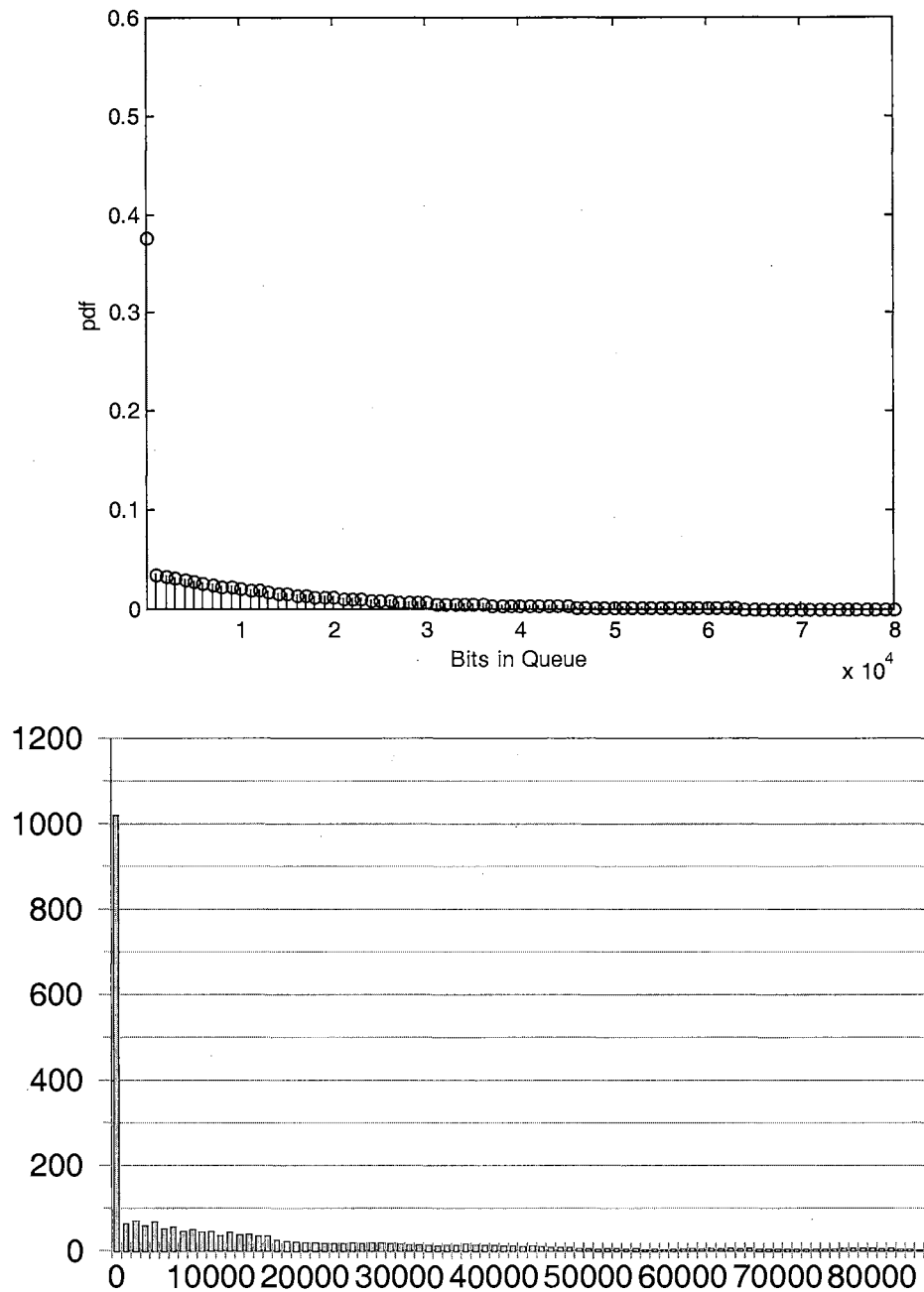


Figure VI-6: PDF of queue size (for 1 frame/packet, output line speed of 20 kbps, trunk load of 61%) from this work (top) and OPNET simulation (bottom)

For a load of 81% and output line speed of 15 kbps, as shown in Figure VI-7, the queue is empty 20% of the time for this work in comparison to 20% for OPNET.

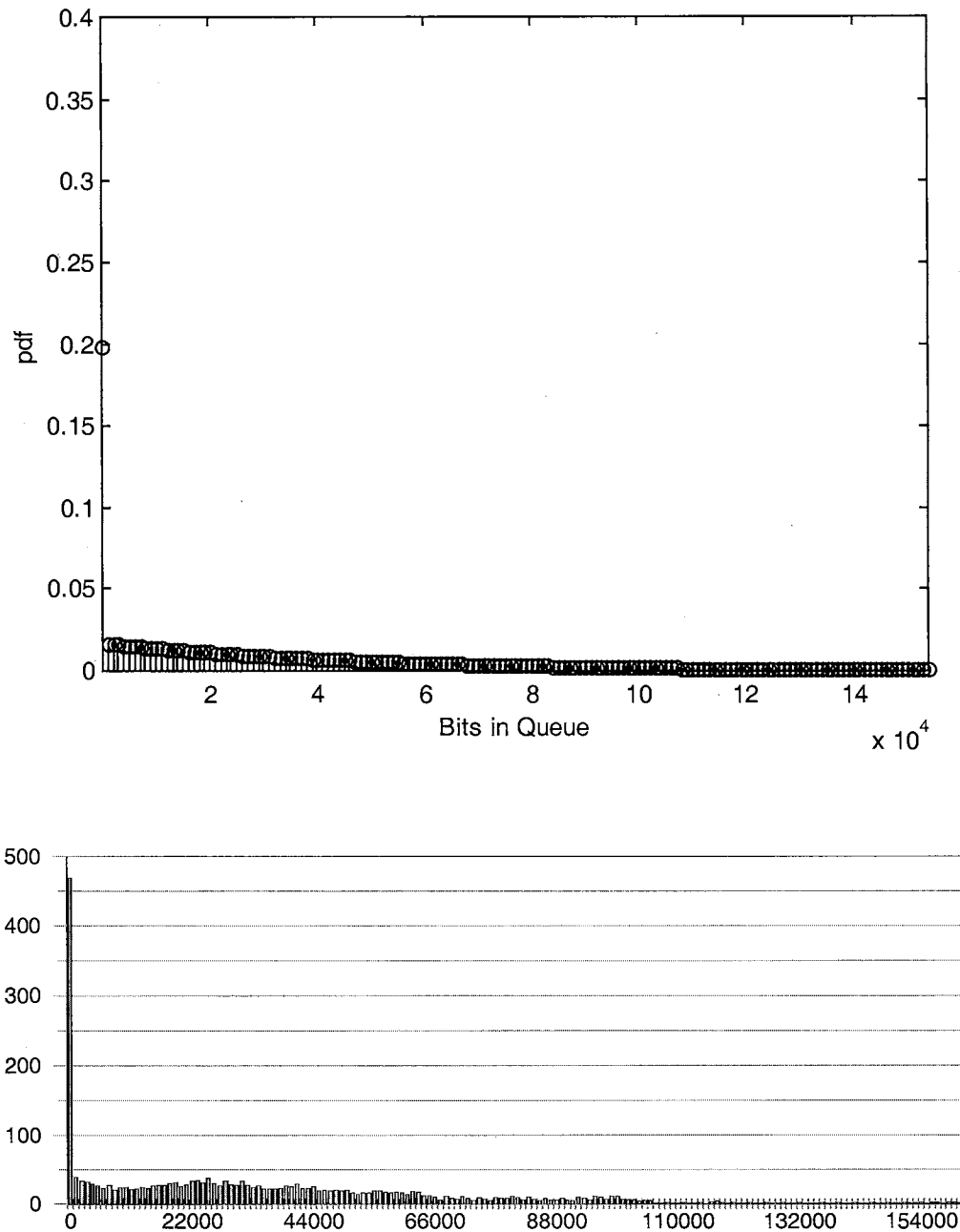


Figure VI-7: PDF of queue size (for 1 frame/packet, output line speed of 15 kbps, trunk load of 81%) from this work (top) and OPNET simulation (bottom)

Figure VI-8 shows the queue PDF for 4 frames/packet for $S_o = 8.8$ kbps, line load of 61%. The other parameters used are $T = 40$ ms, $n_f = 4$, $t_f = 10$ ms, $P_s = 544$ bits, and $S_{in} = 1.54$ Mbps. The queue is empty 33.9% and 33% for this work and OPNET, respectively.

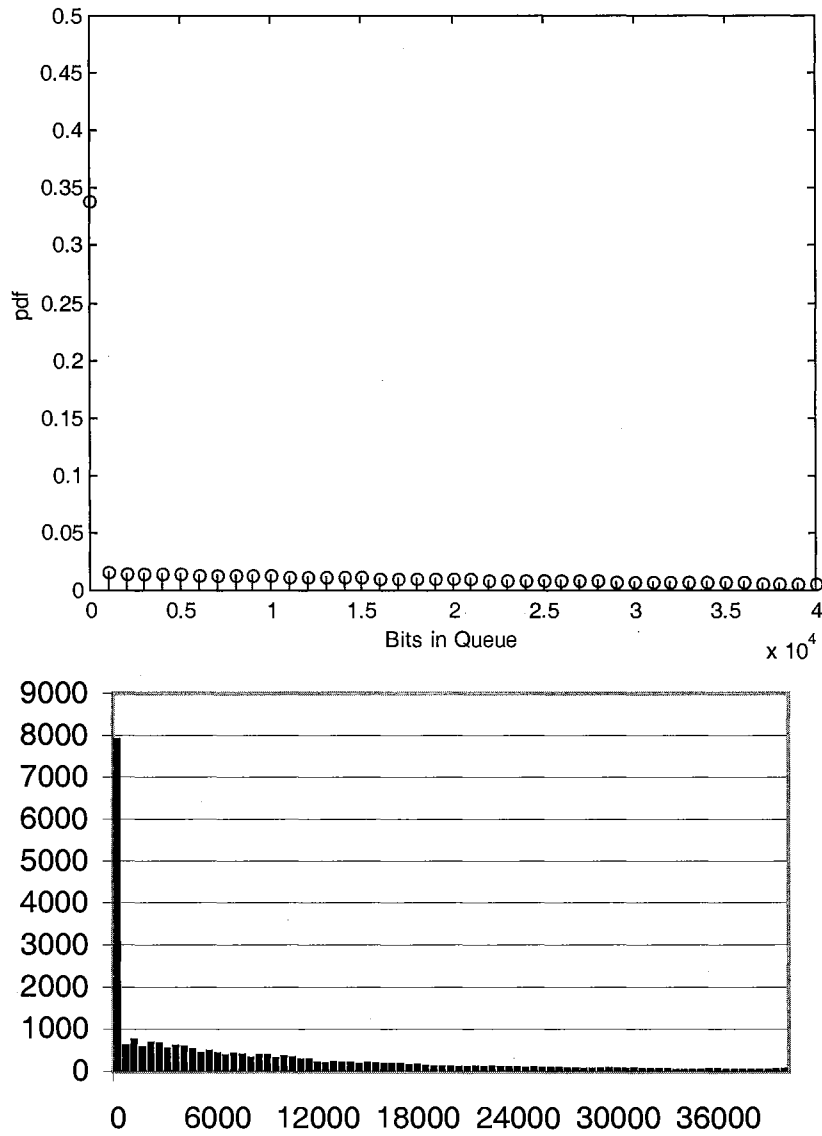


Figure VI-8: PDF of queue size (for 4 frames/packet, output line speed of 8.8 kbps and line load of 61 %) from this work (top) and OPNET simulation (bottom)

Figure VI-9 shows the queue PDF for $S_o = 6.6$ kbps, line load = 81%. The other parameters used are the same as that of Figure VI-8. Here the queue is empty 17.4% of the time for this work and 19% for the OPNET simulation.

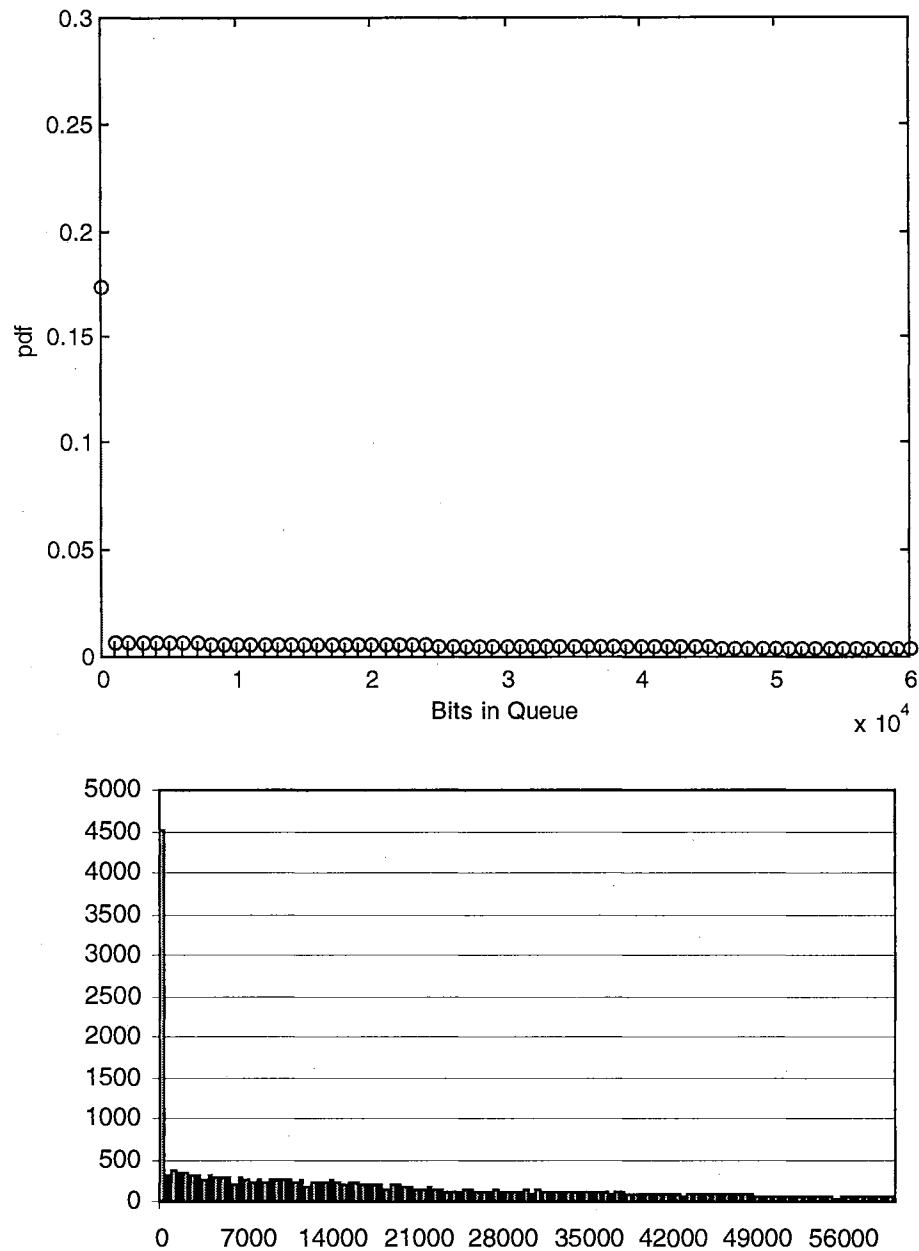


Figure VI-9: PDF of queue size (for 4 frames/packet, output line speed of 6.6 kbps and line load of 81 %) from this work (top) and OPNET simulation (bottom)

A Chi-squared Goodness of Fit test was run to ascertain how well the theoretical PDF matches the OPNET results shown in the figures above. These OPNET plots represent only one experimental run of up to 15 to 30 simulated minutes and, especially under high load conditions, will vary somewhat from experiment to experiment. To help compensate somewhat for this variability, we elected to consolidate the plots into ten bins, effectively averaging and smoothing the experimental outcome. The results of these Goodness of Fit tests are given in the following table.

Line Speed	Chi-squared statistic	$\alpha = 0.01$ Level of Significance	
30 kbps (Figure VI-4)	$\chi_2^2 = 8.497$	$\chi_1^2 = 6.637$	Reject
25 kbps (Figure VI-5)	$\chi_{10}^2 = 5.4425$	$\chi_9^2 = 21.665$	Accept
20 kbps (Figure VI-6)	$\chi_{10}^2 = 15.679$	$\chi_9^2 = 21.665$	Accept
15 kbps (Figure VI-7)	$\chi_{10}^2 = 7.0166$	$\chi_9^2 = 21.665$	Accept
8.8 kbps (Figure VI-8)	$\chi_{10}^2 = 718.749$	$\chi_9^2 = 21.665$	Reject
6.6 kbps (Figure VI-9)	$\chi_{10}^2 = 53.6973$	$\chi_9^2 = 21.665$	Reject

The subscript in the Chi-squared statistic column represents the number of groups of bins, ten in most cases. The subscript in the Level of Significance column represents the number of degrees of freedom. Based on the results, although the theoretical PDF looks quite similar to the experimental results, it cannot be claimed that the theoretical PDF derived in this chapter is always a good statistical match to the experimental outcomes. Likely reasons for these discrepancies include:

*Inaccuracies in the theoretical derivation due to the use of FFT's, and the simplifications used to derive these FFT queue approximations.

*Inaccuracies in the derivation due to the finite number of theoretical PDF's (two-one post talk-spurt PDF and one post silence-interval PDF) used to generate the time average PDF.

*Differences in the internal structure of the switches used in the OPNET simulation and the structure used in the theoretical derivation. The OPNET switches were more complex, having queues at the IP processor and on the output port.

*Additionally, the OPNET simulations in this chapter were run for a short duration of time. A more accurate histogram estimate of the PDF would require a longer run time or an average generated over several runs of short duration.

In conclusion, we observe that the results of the theoretical formulation do not match the simulations results in all the tested cases in a statistical sense. Nevertheless, the results in this chapter visually match the observed simulation results better than any other known VoIP queue derivation. The queue PDF obtained from our theoretical formulation has a large spike at the origin and exhibits similar trends as those obtained from the OPNET simulations.

VII. ANALYSIS OF QUEUE SIZE FOR MULTIPLE VOICE SOURCES

In this Chapter, we attempt to estimate the PDF of the queue size for multiple voice sources. The analysis carried out in Chapter 6 for a single voice source is utilized in part to arrive at results for the multiple voice source case, where N voice sources are connected to the switch as shown in Figure VII-1. The switch has input line speeds of S_{in} and an output line speed of S_o . Three methods were investigated.

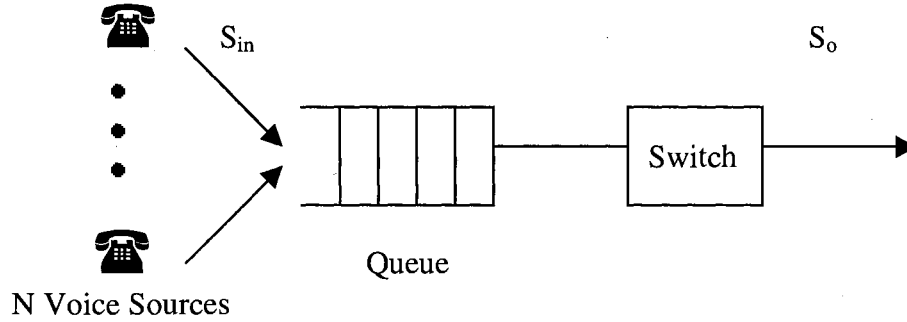


Figure VII-1: Multiple voice source model

Method I

Let X_i be the bits in the queue due to the i^{th} source. Then the resultant bits in the queue due to all N sources are given by a random variable Y where

$$Y = X_1 + X_2 + \dots + X_N \quad (VII-1)$$

If the bits in the queue due to source i are independent of the effects from source j , then the X_i can be considered independent and identically distributed. In this case, the resultant queue PDF $f_Y(y)$ from N sources is given by the N -fold convolution

$$f_Y(y) = f_{X_1} \otimes f_{X_2} \otimes \dots \otimes f_{X_N} \quad (VII-2)$$

In reality the bits in the queue due to source i are, at least to some extent, dependent on source j and their mutual interactions are not easy to quantify.

Method I was investigated under the assumption that, when the f_{X_i} to be used in (VII-2) were generated, the actual output line speed S_o could be modified to a representative value S_o' that properly accounted, on average, for the effect of the other inputs. The hope was that the proper choice of S_o' would make the X_i independent or nearly so. Values that included $S_o' = S_o/N$, and $S_o' = S_o - (N-1)*(\text{Mean Voice Bit Rate})$ were examined. The former equation matches the loading generated by a single source with the load generated on the actual multiple source connection, while the latter subtracted out the bandwidth utilized, on average, by traffic generated by all other sources. All results were disappointing. The contributions to the queue due to the sources are clearly not independent. The results presented in the table below used $S_o' = S_o/N$.

Method II

In the process of investigating Method I it was noted that the PDFs generated for f_{X_i} more closely resembled the OPNET simulation results than the PDF's generated by Method I.

In Method II therefore, we modify the output speed for a single source to match the trunk load of the real case. In other words, the output speed is divided by the number of sources ($S_o' = S_o/N$) and this modified output speed is used to find the queue PDF.

Method III

While the general shape of the PDF from Method II matched OPNET simulations, in general the PDF spike at zero and the PDF spread were off. Method III was based on the hypothesis that matching the load *and* the variance of a single on-off source to the mean and variance of the real multiple source case would generate a PDF estimate that would better match the simulation results and reality. Assuming independent voice conversations on a simplex link, the PDF of the number of sources ON has a binomial distribution. The probability that k out of N sources are ON, P_k , is given by

$$P_k = P(X=k) = \binom{N}{k} \theta^k (1-\theta)^{N-k} \quad (\text{VII-3})$$

where probability that source is ON, $\theta = 0.4$ and N is the total number of voice sources.

The mean of the input bit rate X is

$$\mu = E[X] = N\theta B_p \quad (\text{VII-4})$$

where B_p is the peak bit rate per source. B_p is equal to $P_s/n_f t_f$. P_s , the packet size, is given by equation (VI-3). n_f is the number of frames in a packet and t_f is the frame size. We calculate the second moment of X as

$$E[X^2] = \sum_{k=0}^N P_k (kB_p)^2 \quad (\text{VII-5})$$

where P_k is given in equation (VII-3). Next, we set the mean of the input bit rate of the single source to match the mean input bit rate given by equation (VII-4).

$$0.4 S_{\text{on}} + 0.6 S_{\text{off}} = N\theta B_p \quad (\text{VII-6})$$

where S_{on} and S_{off} are the bit rates when the single source is ON and OFF, respectively.

Then, we set the second moment of X to that of the binomial PDF given by equation (VII-5).

$$0.4 (S_{\text{on}})^2 + 0.6 (S_{\text{off}})^2 = \sum_{k=0}^N P_k (kB_p)^2 \quad (\text{VII-7})$$

We solve equations (VII-6) and (VII-7) to find S_{on} and S_{off} . Using these values we determine the mean bit increase, λ_N^{-1} , and mean bit decrease in the queue, γ_N^{-1} .

$$\lambda_N^{-1} = (S_{\text{on}} - S_o) \alpha^{-1} \quad (\text{VII-8})$$

where $\alpha^{-1} = 1.125$ s is the mean of the talk spurt length and S_o is the output line speed.

The PDF of bit increase in queue is given by

$$f_b(x) = \lambda_N e^{-\lambda_N x}, \quad x \geq 0 \quad (\text{VII-9})$$

The mean bit decrease is

$$\gamma_N^{-1} = (S_o - S_{\text{off}}) \beta^{-1} \quad (\text{VII-10})$$

where $\beta^{-1} = 1.72$ s is the mean of the silence interval. The PDF of bit decrease is given by

$$f_d(x) = \gamma_N e^{\gamma_N x}, \quad x \leq 0 \quad (\text{VII-11})$$

With the PDF's of bit increase, $f_b(x)$, and bit decrease, $f_d(x)$, defined by equations (VII-9) and (VII-11), respectively, we can determine the queue PDF using the approach outlined

in Chapter VI. It should be noted that when S_{on} is less than S_o , the queue remains effectively empty.

Numerical Results

The following table gives the value at which the queue size is, with 99% probability, smaller. Multiple voice sources at different output line speeds and 1 frame per packet were used. Results of Methods I, II and III are compared with those of M/M/1 and OPNET simulation. Note that these Methods all make use of discrete Fast Fourier Transform, and as such have a resolution equal to the bin size- typically a value around 1,000 bits. In the table below, an entry of zero means that the computed PDF lies wholly within the first bin.

Output line speed	# of voice sources	99% queue size (bits)					OPNET Standard Deviation
		Method I	Method II	Method III	M/M/1	OPNET	
30 kbps	2	291,000	206,000	215,000	6,384	211,390	49,133
60 kbps	3	122,000	71,000	53,000	2,736	55,169	17,320
100 kbps	5	163,000	71,000	53,770	2,736	46,227	15,410
200 kbps	7	21,000	8,000	0	1,520	509	8.995
500 kbps	14	0	0	0	1,216	293	0.5
500 kbps	28	811,000	105,000	0	3,344	15,023	6,939
500 kbps	35	1×10^6	252,750	183,300	8,512	179,619	28,998
1.536 Mbps	56	176,000	13,000	0	1,520	331	13

In comparing the results, we find that, generally, Method III is superior to II, which in turn is superior to I. Method I is not accurate showing that the contributions of the sources to the queue are not independent. The PDF shape also tends towards Gaussian, which is not what has been observed in simulations. Method II, where the trunk load is matched, generates queue PDFs that look similar to OPNET's. Figure VII-2 shows a typical PDF generated by this method. Generally, it differs from OPNET simulation results in that the spike at zero is not large enough and the spread of the tail is insufficient. The results of Method III tend to closely match OPNET simulations over the ranges of input and output speeds tested.

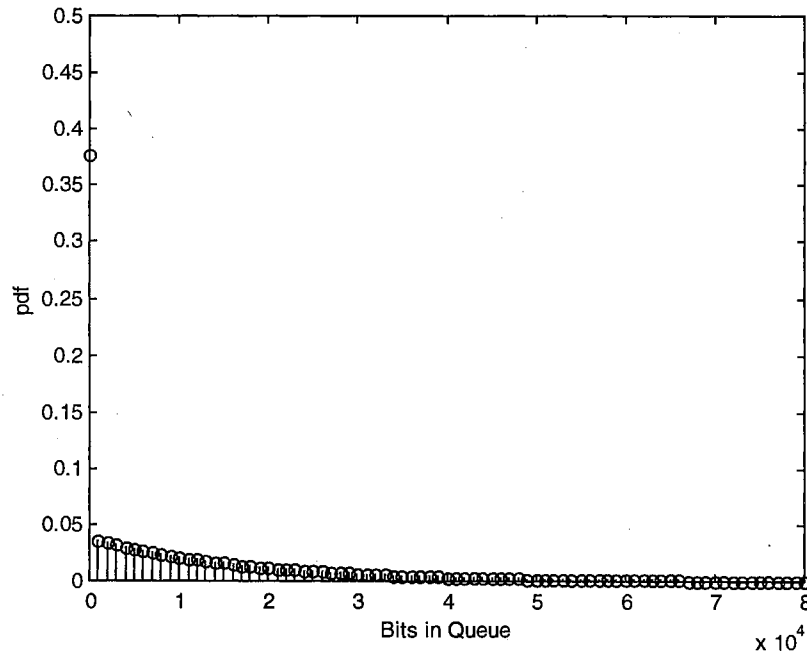


Figure VII-2: PDF of queue size using Method II for 1 frame/packet, output line speed of 60 kbps and 3 voice sources

The PDF of the queue size $f_Y(y)$ can be converted into a delay PDF. This is carried out by scaling the x-axis by $(1/S_0)$. The analysis carried out in Chapter IV

provided us with average delay, that is 50% of the packets can be delivered with in the end-to-end delivery time assuming a symmetrical delay distribution. From the delay PDF computed in this chapter, we can estimate the load such that the 99% of the packets can be delivered in time, at least for instances of a small number of voice sources.

For example, suppose $S_o = 60$ kbps and 3 voice sources are connected to a switch, the load is 61% and it has been found that there is approximately a 99% probability that there are less than or equal to 53,000 bits in the queue. This results in a 99% probability that the delay seen by a packet is, at most, 883 ms. Equation IV-2 gives an average delay of 335 ms for a 61% load, $S_o = 60$ kbps and 1 frame/packet. A large percentage of these packets would require a greater time than the average to traverse the switch. From the delay PDF, were it necessary to get 99% of the packets across the switch with a delay of at most 335 ms, there can be only 2 sources. Hence, the load would have to be reduced from 61% to 41%. The queue PDF for 2 sources, $S_o = 100$ kbps and 1 frame/ packet, has a spike of almost 1 at $x = 0$. This means that the queue and the delay are very small.

VIII. SUMMARY AND CONCLUSIONS

This work is concerned with the analysis of voice over IP (VoIP), which is expected to gain widespread usage in the telecommunication systems of the future. In reviewing the current literature, we have found that little or no work has been carried out for a generalized model of the queue size distribution over a packet switched transmission system carrying pure VoIP traffic. Previous work has either been focused on voice over ATM or the modeling of delay distributions using M/M/1 or M/D/1 queuing models. These models do not accurately represent the arrival and service processes of voice traffic over a packet network. In this dissertation, we have developed a generalized model of the queue distribution of real-time voice over IP for a single voice source and then extended it to account for multiple voice sources. This analysis has been carried out using the most recently introduced G. 729 series voice coders, but is general enough so that it can be extended to other variable rate coders.

In Chapter II, we have provided an overview of the related work carried out by other researchers in the area of VoIP. Starting with the initial work carried out at Bell Labs by Paul Brady, who developed the speech model utilized in this work, we have followed the research in the transmission of voice using ATM technology. The focus of much of the research work in VoIP has been done using M/M/1 or M/D/1 queuing models. Most recently, a group of researchers have provided results on average delay on various types of networks.

In Chapter III, we have determined the end-to-end delay for VoIP networks based on absolute end-to-end delivery, using a fixed rate coder. The coder used is ITU-T

Recommendation G.729 Annex A, which has a voice frame size of 10 ms and bit-rate of 8 kbps. Voice sources here generate traffic in a deterministic manner for the duration of the conversation. Numerical results presented include the number of calls supported as a function of trunk load, and delay vs. number of routers for various loads. These results allow us to determine the number of calls that can be supported for a given load and a fixed end-to-end delay of 150 ms. In addition, we have determined the maximum number of frames that can be put in a packet for a fixed load and delay. Numerical results show that for a fixed rate coder and trunk loads ranging up to 90%, we cannot put more than 3 voice frames per packet for 4 hops and expect to meet the end-to-end delay criterion.

In Chapter IV, we have carried out similar analysis for a variable rate VoIP coder, ITU-T Recommendation G.729 Annex B, which generates traffic in a random manner, outputting little or no traffic during pauses in the conversation. Self-similar traffic queuing theory was used to estimate queuing delays, with a Hurst parameter, $H = 0.85$ [98], a value in the range of that found for actual Internet traffic. Numerical results presented in this Chapter include the number of calls supported as a function of backbone trunk load, and delay vs. number of routers for various loads. This analysis is, however, based on averages, and as such a considerable number of packets would arrive with end-to-end delays greater than the target used in this dissertation of 150 ms. A knowledge of the packet delay or queue distribution at switches is required for an improved analysis.

Chapter V deals with the analysis of the voice packet size. We have based our work on the speech activity model developed at Bell Labs by Paul Brady [11]. The cumulative distribution function (CDF) of the talk spurt and pause from Brady's model is used to determine the probability density function (PDF) of the talk spurt and pause. The

resulting PDF of the talk spurt is used to arrive at a generalized expression for the PDF of the number of voice packets in a talk spurt and the packet size for fixed rate and variable rate coders. From the numerical results, we have found that the probability of getting a full packet (4 frames/packet) is 0.9737 whereas the probability of getting 1, 2, or 3 frames in a packet is 0.0088 for a variable rate coder. In case of the fixed rate coder, the probability of getting a full packet (4 frames of voice in a packet) is 0.3895 and the probability of getting an empty (no voice) packet is 0.5896. The probability of getting 1, 2 or 3 frames of voice in a packet is $\cong 0.007$.

In Chapter VI, we have used the PDF of talk spurt and pause, and converted them into PDF of bit increase and bit decrease in queue, respectively. The resulting PDF's are then utilized in the computation of the PDF of the queue size for a single voice source model. The computed results for 1 frame/packet and various output line speeds are compared with those obtained from OPNET simulation. The computed results compared well with OPNET simulation. We have also provided the queue size PDF for 4 frames/packet.

The analysis carried out in Chapter VI for a single voice source has been extended to accommodate multiple sources in Chapter VII. The queue PDF is obtained using three different methods. Numerical results presented in this Chapter include bits in queue when the area under queue size PDF approaches 99% for 1 frame/packet for various output speeds. Results here agree reasonably well with OPNET simulations.

Finally, we present some recommendation for future research. In this work we considered that we only have voice traffic. Therefore, an obvious extension to this work could be to determine the queue size PDF for voice packets when the traffic contains both

voice and data, with and without the use of priorities. Also in this work we considered the network to have a single switch, with voice sources connected directly to the switch. Modifications to queue size PDFs which account for the correlation that occurs as traffic is passed through multiple switches could be another area of future research.

While the results of Chapter 7 show that Method III matches OPNET simulations fairly well over the range of parameters tested, it cannot be categorically stated that Method III will work under all choices of input/output speeds. No proof or derivation has been provided to show that this technique will always work. Hence another area of research could be to explore the workings of, and possibly modify, the multiple source model at high speeds, in the upper Mbps and lower Gbps range.

To be more useful for a carrier, a method can be developed that, given the tolerable delay through a switch, yields the load that the switch can support. In this work, the techniques we have developed do the opposite. That is, for a given load we can determine the delay distribution. It would be useful to generate a set of 'load' versus 'percentage delay' curves, or formulae, which would allow carrier personnel to quickly estimate the tolerable load given some allowable delay through a switch.

REFERENCES:

- [1] ITU-T Recommendation H.323, "Packet based multimedia communication systems," 1998.
- [2] RFC793, "Transmission control protocols," 1981.
- [3] RFC791, "Internet protocol," 1981.
- [4] RFC1889, "RTP profile for audio and video conferences with minimal control," 1996.
- [5] ITU-T Recommendation G.723.1, "Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3kbits/s," 1996.
- [6] ITU-T Recommendation G.728, "Coding of speech at 16kbits/s using low-delay code excited linear prediction," 1992.
- [7] ITU-T Recommendation G.729, "Coding of speech at 8kbits/s using conjugate-structure algebraic-code-excited linear-prediction," 1996.
- [8] ITU-T Recommendation G.711, "Pulse code modulation (PCM) of voice frequencies," 1996.
- [9] ITU-T Recommendation G.729, Annex A, "Reduced complexity 8kbits/s CS-ACELP speech coder," 1996.
- [10] ITU-T Recommendation G.729, Annex B, "A silence compression scheme for G.729 optimized for terminals conforming to ITU-T V.70".
- [11] P. T. Brady, "A model for on-off speech patterns in two-way conversation," The Bell System Technical Journal, vol. 48, no. 7, pp. 2445-2472, September 1969.
- [12] R. C. Cox, "Three new speech coders from the ITU cover a range of applications," IEEE Comm. Mag., September 1997, pp. 40-47.
- [13] M. E. Perkins, K. Evans and L. A. Thorpe, "Characterizing the Subjective Performance of the ITU-T 8 kb/s Speech Coding Algorithm-ITU-T G.729," IEEE Comm. Mag., September 1997. pp. 74-81.
- [14] R. Salami, C. Laflamme, B. Bessette and J-P Adoul, "ITU-T G.729 Annex A: Reduced Complexity 8 kb/s CS-ACELP Codec for Digital Simultaneous Voice and Data," IEEE Comm. Mag., September 1997, pp. 56-63.

- [15] A. Benyassine, E. Shlomot and H. Su, "ITU-T Recommendation G.729, Annex B: A silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications," *IEEE Comm. Mag.* September 1997, pp. 64-72.
- [16] H. Miedema and M. G. Schachtman, "TASI quality – effect of speech detectors and interpolation," *The Bell System Technology Journal*, pp. 1455-1473, July 1962.
- [17] J. M. Fraser, D. B. Bullock and N. G. Long, "Over-all characteristics of a TASI system," *The Bell System Technology Journal*, pp. 1439-1454, July 1962.
- [18] J. J. Dubnowski and R. E. Crochiere, "Variable rate coding of speech," *The Bell System Technology Journal*, vol. 58, no.3, pp. 577-600, March 1979.
- [19] F. Babich, E. Valentinuzzi and F. Vatta, "Transmission of multimode and variable-rate encoded speech samples on packet switched radio networks handling wide band voice information," *Proceedings of the IEEE 1997 International Symposium on Personal, Indoor and Mobile Radio Communications, Part 3 (of 3)*, vol. 3, pp. 913-917, Helsinki, Finland, 1997.
- [20] S. Nanda, D. J. Goodman and U. Timor, "Performance of PRMA: a packet voice protocol for cellular systems," *IEEE Transactions on Vehicular Technology*, vol. 40, pp. 584-598, August 1991.
- [21] B. Tsybakov and N. Georganas, "On self-similar traffic in ATM queues: definitions, overflow probability bound, and cell delay distribution," *IEEE/ACM Transaction on Networking*, vol.5, no. 3, pp. 397-409, June 1997.
- [22] B. Li and X. Cao, "Experimental results on the impact of cell delay variation on speech quality in ATM networks," *Proceedings of IEEE ICC*, vol. 1, pp. 477-481, 1998.
- [23] F. Hao, I. Nikolaidis and E. Zegura, "Efficient simulation of ATM networks with accurate end-to-end delay statistics," *Proceedings of IEEE ICC*, vol. 3, pp. 1799-1804, 1998.
- [24] Y. H. Kim and C. K. Un, "Performance analysis of statistical multiplexing for heterogeneous bursty traffic in an ATM network," *IEEE Transactions on Communication*, vol. 42, pp. 745-753, February/March/April 1994.
- [25] M. A. Saleh, I. W. Habib and T. N. Saadawi, "Simulation Analysis of a Communication Link with Statistically Multiplexed Bursty Voice Sources," *IEEE J. Selected Areas in Comm.*, vol. 11, no.3, April 1993, pp. 432-441.

- [26] B. Maglaris, D. Anastassiou, P. Sen, G. Karlsson and J. D. Robbins, "Performance models of statistical multiplexing in packet video communications," *IEEE Transactions on Communications*, vol. 36, no. 7, pp. 834-844, July 1988.
- [27] P. Sen, B. Maglaris, N. Rikli and D. Anastassiou, "Models for packet switching of variable-bit-rate video sources," *IEEE Journal on Selected Areas in Communications*, vol. 7, no. 5, pp. 865-869, June 1989.
- [28] M. Nomura, T. Fuji and N. Ohta, "Basic characteristics of variable rate video coding in ATM environment," *IEEE Journal on Selected Areas in Communication*, vol. 7, no. 5, pp. 752-760, June 1989.
- [29] W. Verbiest and L. Pinnoo, "Variable bit rate video codec for asynchronous transfer mode networks," *IEEE Transactions on Communication*, vol. 7, pp. 761-770, June 1989.
- [30] N. Yin and M. G. Hluchyj, "Simple models for statistically multiplexed data traffic in cell relay networks," *Proceedings of IEEE GLOBECOM '93*, pp. 824-829, December 1993.
- [31] A. Elwalid and D. Mitra, "Fluid models for the analysis and design of statistical multiplexing with loss priorities on multiple classes of bursty traffic," *Proceedings of IEEE INFOCOM '92*, vol. 1, pp. 415-425, 1992.
- [32] T. Hou and A. Wong, "Queuing analysis for ATM switching of mixed continuous-bit-rate and bursty traffic," *Proceedings of IEEE INFOCOM '90*, pp. 660-667, 1990.
- [33] J. W. Forgie and A. Nemeth, "An efficient packetized voice/data network using statistical flow control," *Proceedings of International Conference on Communication*, vol. 3, pp. 44-48, June 1977.
- [34] L. Jacob and A. Kumar, "Delay performance of some scheduling strategies in an input queuing ATM switch with multiclass bursty traffic," *IEEE/ACM Transactions on Networking*, vol. 4, no. 2, pp. 258-271, April 1996.
- [35] K. Kondo and M. Ohno, "Packet speech transmission on ATM networks using a Variable rate embedded ADPCM coding scheme," *IEEE Transactions on Communication*, vol. 42, pp. 243-247, February/March/April 1994.
- [36] K. Sato, H. Nakada and Y. Sato, "Variable rate speech coding and network delay analysis for universal transport network," *IEEE INFOCOM Proceedings*, pp. 771-780, 1988.

- [37] F. Beritelli, A. Lombardo, S. Palazzo and G. Schembra, "Performance analysis of an ATM multiplexer loaded with VBR traffic generated by multimode speech coders," *IEEE Journal on Selected Areas in Communication*, vol. 17, pp. 63-79, January 1999.
- [38] K. Chandra and A. Reibman, "Modeling traffic and statistical gains for multimedia applications," *Proceedings of International Workshop on Community Networking*, pp. 171-178, 1995.
- [39] S. J. Golestani, "Duration-limited statistical multiplexing of delay-sensitive traffic in packet networks," *IEEE INFOCOM Proceedings*, pp. 323-332, 1991.
- [40] F. Ishizaki, T. Takine and Y. Oie, "Delay analysis for real-time and non real-time traffic streams under a priority cell scheduling," *IEEE GLOBECOM Proceedings*, vol. 5, pp. 3007-3012, 1998.
- [41] B. Steyaert and H. Bruneel, "An effective algorithm to calculate the distribution of the buffer contents and the packet delay in a multiplexer with bursty sources," *IEEE GLOBECOM '91 Proceedings*, vol. 1, pp. 471-475, 1991.
- [42] G. Nong, M. Hamdi and K. Letaief, "Efficient scheduling of variable-length IP packets on high-speed switches," *IEEE GLOBECOM '99*, pp. 1407-1411, 1999.
- [43] J. Schormans, J. Pitts, R. Mondragon, E. Scharf, A. Pearmain and C. Phillips, "Design rules for buffering overlapping Pareto processes in packetized networks," *Electronics Letters*, vol. 36, pp. 1086-1088, June 2000.
- [44] J. Schormans and J. Pitts, "Decay rate (ER) modelling of G/D/1 queue, and results for ATM telecommunication," *Electronics Letters*, vol. 34, pp. 943-947, May 1998.
- [45] C. Weinstein and J. Forgie, "Experience with speech communication in packet networks," *IEEE Journal on Selected Areas in Communication*, vol. SAC-1, no. 6, pp. 963-980, December 1983.
- [46] M. Baldi and F. Risso, "Efficiency of packet voice with deterministic delay," *IEEE Communication Magazine*, pp. 170-177, May 2000.
- [47] S. Li, "Generating function approach for discrete queuing analysis with decomposable arrival and service Markov Chains," *Proceedings of IEEE INFOCOM*, pp. 2168-2177, 1992.
- [48] C. Yuan and J. Silvester, "Queuing analysis of delay constrained voice traffic in a packet switching system," *IEEE Journal on Selected Areas in Communication*, vol. 7, no. 5, pp. 729-738, June 1989.

- [49] S. Dravida and K. Sriram, "End-to-end performance models for variable bit rate voice over tandem links in packet networks," *IEEE Journal on Selected Areas in Communication*, vol. 7, no. 5, pp. 718-727, June 1989.
- [50] Y. C. Jenq, "Approximations for packetized voice traffic in statistical multiplexer," *Proceedings of IEEE INFOCOM*, pp. 256-259, 1984.
- [51] T. Bially, B. Gold and S. Seneff, "Technique for adaptive voice flow control in integrated packet networks," *IEEE Transactions on Communication*, vol. Com-28, no. 3, pp. 325-333, March 1980.
- [52] C. Weinstein, "Fractional speech loss and talker activity model for TASI and for packet-switched speech," *IEEE Transactions on Communication*, vol. Com-26, no. 8, pp. 1253-1257, August 1978.
- [53] K. M. Elsayed and H. G. Perros, "Statistical multiplexing with arbitrary on/off sources," *Proceedings of IEEE Conference on Computers and Communications*, pp. 487-493, 1996.
- [54] S. Ganguly and T. Stern, "Performance evaluation of a packet voice system," *IEEE Transactions on Communications*, vol. 37, no. 12, pp. 1394-1397, December 1989.
- [55] R. Tucker, "Accurate method for analysis of a packet-speech multiplexer with limited delay," *IEEE Transactions on Communications*, vol. 36, no. 42, pp. 479-483, April 1988.
- [56] M. Borella and G. Brewster, "Measurement and analysis of long-range dependent behavior of Internet," *IEEE INFOCOM Proceedings*, pp. 497-504, 1998.
- [57] Q. Li and D. Mills, "On the long-range dependence of packet round-trip delays in Internet," *Proceedings of IEEE International Conference on Communications*, pp. 1185-1191, 1998.
- [58] V. Paxson and S. Floyd, "Wide area traffic: the failure of Poisson modeling," *IEEE/ACM Transactions on Networking*, vol. 3, no. 3, pp. 226-244, June 2000.
- [59] F. Yegenoglu, F. Faris and O. Qadan, "A model for representing wide area Internet packet behavior," *IEEE*, pp. 167-173, 2000.
- [60] D. Sanghi, A. Agrawala, O. Gudmundsson and B. Jain, "Experimental assessment of end-to-end behavior on Internet," *IEEE INFOCOM Proceedings*, vol. 2, pp. 867-874, 1993.
- [61] J. Bolot, "End-to-end packet delay and loss behavior in the Internet," *Computer Communication Review*, vol. 23, no. 4, pp. 289-298, 1993.

- [62] M. Lucas, D. Wrege, B. Dempsey and A. Weaver, "Statistical characterization of wide-area IP traffic," *Proceedings of IEEE International Conference on Computer Communications and Networks*, pp. 442-447, 1997.
- [63] V. Paxson, "End-to-end Internet packet dynamics," *IEEE/ACM Transactions on Networking*, vol. 7, no. 3, pp. 277-292, June 1999.
- [64] K. Thompson, G. Miller and R. Wilder, "Wide-area Internet traffic patterns and characteristics," *IEEE Network*, pp. 10-23, November/December 1997.
- [65] J. Andren, M. Hilding and D. Veitch, "Understanding end-to-end Internet traffic dynamics," *IEEE GLOBECOM*, vol. 2, pp. 1118-1122, 1998.
- [66] E. Tse-Au and P. Morreale, "End-to-end QoS measurement: analytic methodology of application response time vs. tunable latency in IP networks," *Proceedings of IEEE NOMS*, pp. 129-142, 2000.
- [67] T. Elteto and S. Molnar, "On the distribution of round-trip delays in TCP/IP networks," *IEEE Conference on Local Computer Networks*, pp. 171-181, 1999.
- [68] H. Tang, S. Huang and H. Chen, "Internet flow blocking probability calculation," *Canadian Conference on Electrical and computer Engineering*, pp. 659-663, 2000.
- [69] S. Golestani, "Network delay analysis of a class of fair queuing algorithms," *IEEE Journal on Selected Areas in Communication*, vol.13, no.6, pp. 1057-1070, August 1995.
- [70] V. Ribeiro, R. Riedi, M. Crouse and R. Baraniuk, "Multiscale queuing analysis of long-range-dependent network traffic," *IEEE INFOCOM*, pp. 1026-1035, 2000.
- [71] M. Girish and J. Hu, "Modeling of correlated arrival processes in the Internet," *Proceedings of IEEE Conference on Decision and Control*, pp. 4454-4459, December 1999.
- [72] C. Casetti and M. Meo, "A new approach to model the stationary behavior of TCP connections," *IEEE INFOCOM*, pp. 367-375, 2000.
- [73] J. Muppala, T. Banerjee and A. Tyagi, "VoIP performance on differentiated services enabled network," *Proceedings of IEEE International Conference on Networks*, pp. 419-423, 2000.
- [74] V. Abreu-Sernandez and C. Garcia-Mateo, "Adaptive multi-rate speech coder for VoIP transmission," *Electronics Letters*, vol. 36, no. 23, pp. 1978-1979, November 2000.

- [75] T Miyata, H. Fukuda and S. Ono, "New network QoS measures for FEC-based Audio applications on the Internet," IEEE International Performance, Computing and Communication, pp. 355-362. 1998.
- [76] M. Baker, "Speech transport for packet telephony and voice over IP," Proceedings of SPIE Conference on Internet II, vol. 3842, pp. 242-251, September 1999.
- [77] T. Kostas, M. Borella, I. Sidhu, G. Schuster, J. Grabiec and J. Mahler, "Real-time voice over packet-switched networks," IEEE Networks, pp. 18-27, January/February 1998.
- [78] P. Mishra and H. Saran, "Capacity management and routing policies for voice over IP traffic," IEEE Networks, pp. 20-27, March/April 2000.
- [79] T. Hoshi, K. Tanigawa and K. Tsukada, "Voice stream multiplexing between IP telephony gateways," IEICE Transactions on Information and System, vol. E82-D, no. 4, pp. 838-845, April 1999.
- [80] B. Goodman, "Internet telephony and modem delay," IEEE Network, pp. 8-16, May/June 1999.
- [81] M. Hamdi, O. Vercheure and J. Hubaux, "Voice service interworking for PSTN and IP networks," IEEE Communications Magazine, pp. 104-111, May 1999.
- [82] D. Vleeschauwer, G. Petit, B. Steyaert, S. Wittevrongel and H. Bruneel, "An accurate closed-form formula to calculate the dejittering delay in packetized voice transport," Proceedings of the IFIP-TC6 / European Commission International Conference Networking, pp. 374-385, May 2000.
- [83] D. Vleeschauwer, J. Janssen and G. Petit, "Delay bounds for low bit rate voice transport over IP networks," Proceedings of the SPIE Conference on Performance and Control of Network Systems III, vol. 3841, pp. 40-48, September 1999.
- [84] F. Poppe, D. Vleeschauwer and G. Petit, "Choosing the UMTS air interface parameters, the voice packet size and the dejittering delay for a Voice-over-IP call between a UMTS and a PSTN party," IEEE INFOCOM, pp. 805-814, 2001.
- [85] D. Vleeschauwer, J. Janssen and G. Petit, "Voice over IP in Access networks," Proceedings of the 7th IFIP Workshop on Performance Modelling and Evaluation of ATM/IP Networks, June 1999.
- [86] J. Janssen, R. Windey, D. Vleeschauwer, G. Petit and J. Leroy, "Maximum delay bounds for voice transport over satellite internet access networks," Proceedings of the 4th IEEE International Workshop on Satellite-Based Information Services, pp. 48-55, December 1999.

- [87] F. Poppe, D. Vleeschauwer and G. Petit, "Guaranteeing quality of service to packetized voice over the UMTS air interface," Proceedings of the Eight International Workshop on Quality-of-service, June 2000.
- [88] D. Vleeschauwer, J. Janssen, E. Desmet and G. Petit, "Tolerable delay bounds for low bit rate voice transport," Proceedings of WTC/ISS 2000, May 2000.
- [89] M. Buchli, D. Vleeschauwer, J. Janssen, A. Moffaert and G. Petit, "On the efficiency of voice over integrated services using guaranteed service," Proceedings of the 2nd IP-Telephony Workshop, pp. 6-14, April 2001.
- [90] J. Janssen, D. Vleeschauwer and G. Petit, "Delay and distortion bounds for packetized voice calls of traditional PSTN quality," Proceedings of the 1st IP-Telephony Workshop, pp. 105-110, April 2000.
- [91] J. Schormans, J. Pitts, E. Scharf, A. Pearmain and C. Phillips, "Buffer overflow probability for multiplexed on-off VoIP sources," Electronics Letters, vol. 36, no. 6, pp. 523-524, March 2000.
- [92] R. Stewart, J. Schormans, J. Pitts, E. Scharf, A. Pearmain and C. Phillips, "Analysis of local internet/enterprise networking multiplexer for ON-OFF VoIP sources," Electronics Letters, vol. 36, no. 21, pp. 1825-1826, October 2000.
- [93] D. Minoli and E. Minoli, Delivering Voice over IP Networks, John Wiley and Sons, New York, 1998.
- [94] P. Brady, "A technique for investigating on-off patterns of speech," BSTJ, pp. 1-22, January 1965.
- [95] P. Brady, "A statistical analysis of on-off patterns in 16 conversations," BSTJ, pp. 73-91, January 1968.
- [96] ITU-T Recommendation G.114 "One-way transmission time," February 1996.
- [97] P. Goyal, A. Greenberg, C. Kalmanek, W. Marshall, P. Mishra, D. Nortz and K. Ramakrishnan, "Integration of call signaling and resource management for IP telephony," IEEE Network, pp. 24-32, May/June 1999.
- [98] M. Roughan, D. Veitch and P. Abry, "Real-time estimation of the parameters of long-range dependence," IEEE/ACM Transactions on Networking, vol. 8, no. 4, pp. 467-478, August 2000.
- [99] W. Stallings, High Speed Networks TCP/IP and ATM Design Principles, Prentice Hall, 1998.

- [100] H. Heffes and D. Lucantoni, "A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance," IEEE Journal on Selected Areas in Communication, vol. SAC-4, no.4, pp. 856-867, September 1986.

VITA 2

Ritu Singh

Candidate for the Degree of

Doctor of Philosophy

Thesis: QUEUE DISTRIBUTION OF REAL TIME TRANSPORTATION OF VOICE
OVER IP

Major Field: Electrical Engineering

Biographical:

Personal Data: Born in Rohtak, India, On April 25,1961, the daughter of Rattan Singh and Shakuntla Sandhu.

Education: Graduated from St. Theresa School in April 1977; received Bachelor of Arts degree in English Literature from Kurukshetra University in April 1980; received Bachelor of Science and Master of Science in Electrical Engineering from The University of Tulsa in May 1990 and May 1996, respectively. Completed the requirements for the Doctor of Philosophy degree with a major in Electrical Engineering at Oklahoma State University in August 2002.

Experience: Served as graduate research assistant at The University of Tulsa, 1990 to 1993; research assistant at Oklahoma State University, 1998 to 2001.

Professional Memberships: Phi Kappa Phi, Phi Gamma Kappa, Eta Kappa Nu and Tau Beta Pi.